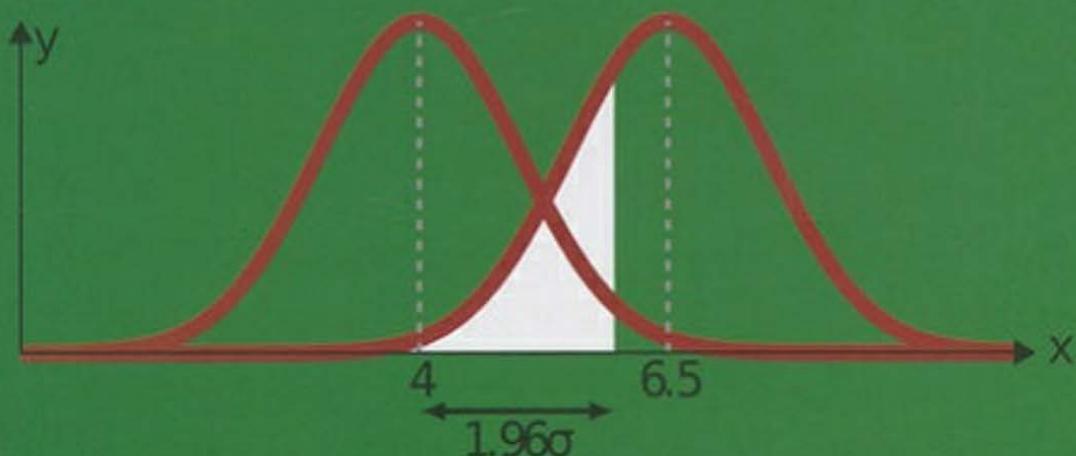


Luis Eduardo Girón Cruz

Econometría Aplicada

Usando Stata 13



Luis Eduardo Girón Cruz

**ECONOMETRÍA APLICADA
Usando Stata 13**



Girón Cruz, Luis Eduardo, 1959-

Econometría aplicada : Usando Stata 13 / Luis Eduardo Girón Cruz – Santiago de Cali : Pontificia Universidad Javeriana, Sello Editorial Javeriano, 2017.

125 páginas; ilustraciones; 27cm.

Incluye referencias bibliográficas.

ISBN: 978-958-8856-97-1

1. Econometría – Manuales 2. Análisis de regresión 3. Stata (Programa para computador) 4. Estadística – Programas para computador 5. Estadística – Procesamiento de datos I. Girón Cruz, Luis Eduardo II. Pontificia Universidad Javeriana (Cali). Facultad de Ciencias Económicas y Administrativas. Departamento de Economía.

SCDD 330.015195 ed. 23 CO-CaPUJ
malc/17



Pontificia Universidad
JAVERIANA
Cali

ECONOMETRÍA APLICADA

Usando Stata 13

Luis Eduardo Girón Cruz, EDITOR O COMPILADOR

ISBN: 978-958-8856-97-1

Primera edición: agosto 2017

Pontificia Universidad Javeriana Cali

Rector: P. Luis Felipe Gómez Restrepo, S.J.

Vicerrectora Académica: Ana Milena Yoshioka Vargas

Vicerrector del Medio Universitario: Libardo Valderrama Centeno S.J.

Decano de la Facultad de Ciencias Económicas y Administrativas: Alberto Arias Sandoval

Departamento de Economía: Luis Fernando Aguado Quintero

Coordinadora Sello Editorial Javeriano Cali

Claudia Lorena González González

e-mail: Cl大陆.gonzalez@javerianacali.edu.co

Diseño y Concepto gráfico: Juan Pablo Girón

Desarrollo eBook: Lápiz Blanco S.A.S.

©Derechos Reservados Traducción

©Sello Editorial Javeriano

Correspondencia, suscripciones y solicitudes de canje:

Pontificia Universidad Javeriana Cali

Calle 18 No. 118-250, Vía Pance

Teléfonos (57-2) 3218200 Ext: 8489

Santiago de Cali, Valle del Cauca

Agradecimientos

Agradezco primeramente a Dios por darme la oportunidad de culminar este manual.

A Luis Fernando Aguado, director del Departamento de Economía de la Pontificia Universidad Javeriana Cali, por su irrestricto apoyo.

A mi hijo Juan Pablo Girón, por su permanente ayuda en lo referente al diseño a través de programación en L^ATEX.

A Camilo Álvarez, estudiante de octavo semestre del programa de Economía de la Pontificia Universidad Javeriana, por su entusiasmo en la digitación y sus múltiples y valiosas sugerencias.

A todos mis compañeros del Departamento de Economía de la Pontificia Universidad Javeriana, por animarme para la culminación de este manual.

A mis estudiantes de Econometría, por sus sugerencias.

A Andrés Rangel y Geovanny Castro, profesores de Econometría, por sus comentarios en su papel de revisores, que permitieron introducir mejoras a este manual.

A mi esposa e hijos por el apoyo, la paciencia y la comprensión por el tiempo de ellos que invertí en este trabajo.

Presentación

Este manual va dirigido tanto a investigadores como a estudiantes de Economía que adelantan cursos de Econometría básica. Ha sido diseñado con el propósito de que sirva de guía en el manejo del paquete estadístico Stata (StataCorp, 2013).

Si bien existe un sinnúmero de manuales de Stata, este se caracteriza por lo siguiente:

1. Es una guía práctica que explica con un lenguaje sencillo el manejo del Stata, concentrándose en la aplicación de los métodos estadísticos y en la interpretación de las salidas del paquete, más que en los desarrollos teóricos propios de la econometría. Aunque es un manual práctico y de fácil comprensión, no significa que la econometría se reduzca a un clic, ya que el verdadero conocimiento de la econometría está en una sólida fundamentación de la teoría subyacente al tema presentado.
2. Ilustra de forma didáctica, mediante ventanas y/o comandos, cómo obtener los resultados. Además, en algunos capítulos describe la manera de obtener los resultados a través de un archivo .do básico, sin pretender desarrollar programación avanzada en Stata.
3. Sigue de forma secuencial los temas de los cursos de Econometría que se imparten en Economía en las universidades de Colombia. Inicialmente se modela utilizando datos de corte transversal, como lo sugiere (Wooldridge, 2010), para luego pasar a modelar con datos de series de tiempo, ecuaciones simultáneas, modelos de variable dependiente limitada y datos de panel. Un aspecto importante es que se resalta el cuidado que se debe tener al trabajar con datos de series de tiempo, por ello se hace una pequeña introducción al tema de raíces unitarias y cointegración univariada para que el estudiante tenga una visión general del tema, sin pretender profundizar en el mismo, pues este es propio de cursos avanzados de Econometría de las Series de Tiempo.
4. Hace una breve introducción del tema estadístico asociado antes de mostrar la ruta o el comando para obtener un resultado y su interpretación, sin profundizar teóricamente en el mismo, ya que esto corresponde a los textos de Econometría.

La metodología utilizada para que el lector se apropie del manejo del paquete consiste en: (i) desarrollar casos paso a paso, con el objetivo de ilustrar a través de salidas los conceptos teóricos del tema que se está tratando, y (ii) mostrar la ruta por ventanas y comandos para llegar a tales resultados. Al final de algunos capítulos se presenta el programa o archivo do-file, cuya extensión es .do, que es un archivo de texto que contiene o reúne los comandos de Stata utilizados para obtener cada resultado. Estos archivos .do son útiles, pues contienen solo los comandos que han funcionado correctamente y que el estudiante o investigador puede ir copiando para conformar su archivo .do, de tal manera que es posible guardar un trabajo que aún no se ha culminado y retomarlo después sin tener que empezar desde el inicio.

Para ilustrar la ruta de algunos comandos en Stata se usa la siguiente estructura:

Comando

comando 1 → comando 2 → ... → comando n → 

→ es usado para indicar que el usuario tiene que hacer clic con el mouse para ejecutar el comando.

Este manual incorpora ocho capítulos distribuidos de la siguiente manera: en el [capítulo 1](#) se presentan las ventanas que conforman la pantalla de trabajo Stata, la lectura de base de datos, la construcción de base de datos, el análisis de datos y la creación y modificación de variables en la base de datos. En el [capítulo 2](#) se desarrolla la estimación e inferencia del modelo de regresión lineal múltiple en sus diversas formas funcionales, utilizando datos de corte transversal, como lo sugiere (Wooldridge, 2010). En el [capítulo 3](#) se extiende el modelo de regresión múltiple al caso cuando se incorporan variables cualitativas como variables explicatorias. En el [capítulo 4](#) se estudian tres problemas estadísticos, como son la multicolinealidad, la heterocedasticidad y la mala especificación. En el [capítulo 5](#) se desarrolla la estimación e inferencia del modelo de regresión lineal múltiple utilizando datos de series de tiempo; también aspectos como la estabilidad de las estimaciones, las estadísticas de influencia, el problema de autocorrelación, los modelos dinámicos, el test de causalidad de Granger y el concepto de raíces unitarias y cointegración univariada. En el [capítulo 6](#) se tratan las ecuaciones simultáneas y los modelos VAR. En el [capítulo 7](#) se consideran de una manera amplia los modelos de variable dependiente limitada. Para terminar, en el [capítulo 8](#) se hace una breve explicación sobre un tema de interés, como son los datos de panel.

Índice general

1 Introducción al Stata

- 1.1 ¿Qué es el Stata?
- 1.1.1 La pantalla de Stata
- 1.2 Lectura de base de datos
- 1.2.1 Lectura de datos en Excel
- 1.2.2 Lectura de datos .tex
- 1.2.3 Lectura de datos en formato .dta
- 1.3 Construcción de bases de datos
- 1.3.1 Introducción de datos de forma manual
- 1.3.2 Combinación de bases de datos
- 1.4 Análisis de datos
- 1.4.1 Obtención de base datos condensada
- 1.4.2 Análisis descriptivo de datos
- 1.5 Creación y modificación de variables en la base de datos
- 1.6 Apéndices

2 Modelo de regresión lineal múltiple con datos transversales

- 2.1 Estimación por mínimos cuadrados ordinarios (MCO) para datos transversales
- 2.1.1 Explicación de la salida de un modelo de regresión lineal múltiple estimado en Stata
- 2.1.2 Interpretación de los coeficientes de regresión estimados
- 2.1.3 Inferencia en el modelo de regresión lineal múltiple
- 2.1.4 Hipótesis sobre una combinación lineal o conjunto de combinaciones lineales de los coeficientes de regresión
- 2.1.5 Formas funcionales de un modelo

3 Análisis de regresión con información cualitativa

- 3.1 Una variable cualitativa como explicatoria
- 3.1.1 Caso I. La variable cualitativa modifica solo el intercepto
- 3.1.2 Caso II. Cuando la dummy modifica solo la pendiente de años de educación
- 3.1.3 Caso III. Cuando la dummy modifica el intercepto y la pendiente
- 3.2 Más de una variable cualitativa como explicatoria
- 3.2.1 Caso I. Más de una variable cualitativa sin interacción
- 3.2.2 Caso II. Más de una variable cualitativa como explicatoria con interacción

4 Problemas económicos

- 4.1 Multicolinealidad
- 4.1.1 Multicolinealidad perfecta
- 4.1.2 Alta multicolinealidad
- 4.1.3 Pruebas para detectar alta multicolinealidad
- 4.1.4 Medidas remediales a la alta multicolinealidad
- 4.2 Heterocedasticidad
- 4.2.1 ¿Cómo detectar heterocedasticidad?
- 4.2.2 ¿Cómo solucionar heterocedasticidad?
- 4.3 Especificación
- 4.3.1 Pruebas para variables omitidas y forma funcional incorrecta
- 4.3.2 Pruebas para selección de modelos

4.3.3 Criterios de selección de modelos

5 Análisis de regresión con series de tiempo

5.1 Estimación e interpretación de un modelo con datos de series de tiempo

5.1.1 Estadísticas de influencia

5.2 Autocorrelación

5.2.1 ¿Cómo detectar autocorrelación?

5.2.2 ¿Cómo solucionar autocorrelación?

5.3 Modelos econométricos dinámicos

5.3.1 Estimación de un modelo dinámico

5.4 Test de causalidad

5.5 Desestacionalización

5.6 Raíz unitaria y cointegración univariada

5.6.1 Pruebas de raíz unitaria y cointegración univariada

6 Ecuaciones simultáneas

6.1 Modelos VAR

7 Microeconometría

7.1 Modelos de elección discreta binaria

7.1.1 Modelo lineal de probabilidad

7.1.2 El modelo logit

7.1.3 El modelo probit

7.1.4 Modelos de datos de recuento

7.1.5 Modelo de regresión de Poisson

7.1.6 El modelo de regresión Tobit

8 Modelos de regresión con datos de panel

8.1 Modelo de regresión con MCO agrupados o de coeficientes constantes

8.2 Modelo de efectos fijos usando variable dummy

8.3 Modelo de efectos aleatorios (MEFA)

8.4 Estimación de un modelo con datos de panel en Stata

8.4.1 Modelo de panel de coeficientes constantes

8.4.2 Modelo de efectos aleatorios (MEFA)

8.4.3 Modelo de panel de mínimos cuadrados con variable dicótoma de efectos fijos

8.4.4 Test de Hausman para escoger entre efectos fijos y aleatorios

8.5 Apéndice

Referencias

Introducción al Stata

Temas Tratados

- 1.1 ¿Qué es el Stata?**
 - 1.1.1 La pantalla de Stata**
- 1.2 Lectura de base de datos**
 - 1.2.1 Lectura de datos en Excel**
 - 1.2.2 Lectura de datos .tex**
 - 1.2.3 Lectura de datos en formato .dta**
- 1.3 Construcción de bases de datos**
 - 1.3.1 Introducción de datos de forma manual**
 - 1.3.2 Combinación de bases de datos**
- 1.4 Análisis de datos**
 - 1.4.1 Obtención de base datos condensada**
 - 1.4.2 Análisis descriptivo de datos**
- 1.5 Creación y modificación de variables en la base de datos**
- 1.6 Apéndices**

1.1 ¿Qué es el Stata?

Stata es un sistema de análisis estadístico para los profesionales de la investigación. El sitio oficial es <http://www.stata.com/>. Se trata de un entorno propicio para manipular y analizar datos usando métodos estadísticos y gráficos. Stata es un conjunto integrado de módulos que permiten, entre otros, manejo flexible de base de datos; generación o transformación de nuevas variables; almacenamiento de documentos; actualización y edición de archivos; elaboración de gráficos; análisis estadístico descriptivo, inferencial, multivariado; análisis de series de tiempo; y modelación econométrica.

Una característica importante de Stata es que posee un conjunto de bases de datos que pueden ser descargadas y utilizadas para ilustrar el manejo de un comando. La ruta para descargar dichas bases es:

Ventana

```
file→ example Datasets→ ↗ → Example datasets installed with Stata→ ↗
```

En este punto se puede seleccionar tanto el archivo como la descripción de las variables que lo componen. Adicionalmente, para profundizar en las opciones y su sintaxis, el Stata provee la opción **help** de ayuda, la cual permite ampliar el conocimiento de un comando específico. Como Stata se actualiza permanentemente, es posible incorporar las nuevas actualizaciones o comandos que no se encuentren en el programa, esto se logra a través del siguiente comando:

Comando

```
findit nombre del comando
```

findit también permite localizar a través de una palabra clave un comando desconocido.

1.1.1 La pantalla de Stata

La pantalla de Stata está compuesta por seis elementos que se describen a continuación y se muestran en la [Figura 1.1](#).

1. barra de menú.
2. una barra de herramientas.
3. una ventana de resultados.
4. una ventana de comandos.
5. una ventana de revisión.
6. una ventana de variables.

La barra de menú

La barra de menú está compuesta por ocho botones, como se muestra en la [Figura 1.1](#). Al hacer clic (1) en uno de estos botones, Stata muestra un menú desplegable de la lista de comandos disponibles para ese botón. Algunos usuarios prefieren escribir los comandos en la ventana de comandos, mientras que la mayoría prefieren utilizar la interfaz de menús. Sin embargo, para ejecutar grandes programas es recomendable utilizar la ventana de comandos, junto con un archivo de registro (do file).

La barra de herramientas

La barra de herramientas, situada justo debajo de la barra de menús, proporciona un acceso rápido y fácil a muchos procedimientos de uso frecuente, como se muestra en la [Figura 1.2](#).



Figura 1.1: Pantalla del programa Stata.



Figura 1.2: Descripción de la barra de herramientas.

La ventana de resultados

En esta ventana aparecen los resultados de las órdenes ejecutadas. Esta ventana puede despejarse después de realizar estimaciones por medio del comando `cls`. Si se quiere guardar los resultados obtenidos en el análisis para tenerlos disponibles en un archivo de texto, se debe utilizar un archivo con extensión `.log`, el cual realiza este tipo de funciones. La ruta para conformar un archivo `log` es:

Ventana

`file → log → begin → ruta donde se guardará el archivo → nombre del archivo → guardar.`

La ventana de comandos

En esta ventana se escriben las órdenes que se dan a la aplicación.

La ventana de revisión

En esta ventana se listan completamente los comandos ejecutados desde que se inicia la aplicación.

La ventana de variables

En esta ventana se listan las variables contenidas en la base de datos cargada en la aplicación. Si no se tiene ninguna variable, esta ventana aparecerá vacía, como se observa en la [Figura 1.1](#).

1.2 Lectura de base de datos

El Stata permite trabajar los diferentes tipos de datos utilizados en econometría, como son: datos de corte transversal, datos de series de tiempo y datos de panel o longitudinales. En este manual se inicia con modelos que utilizan datos de corte transversal, como lo sugiere (Wooldridge, 2010), para luego pasar a modelos que utilizan datos de series de tiempo y datos de panel. Existen varias formas para leer base de datos en Stata, las más comunes son: Excel, `.tex` y formato `.dta`.

Recomendaciones para la lectura de datos

Antes de presentar las formas más comunes de leer bases de datos en Stata, se deben tener en cuenta las siguientes consideraciones:

- Stata, antes de abrir un nuevo archivo, exige que la ventana de variables esté vacía, esto se logra con el comando `clear`.
- Stata diferencia entre mayúsculas y minúsculas, es por ello que los comandos en Stata deben ser escritos en minúsculas.
- Para guardar un archivo de datos, simplemente se sigue la ruta `file → save as → ruta donde se desea guardar el archivo .dta`.

- Para combinar archivos de datos en Stata, estos se deben haber convertido previamente a formato .dta.

1.2.1 Lectura de datos en Excel

En este caso existen varias posibilidades: la primera, la cual no es muy recomendable, es copiar los datos desde Excel y pegarlos en la hoja de datos proporcionada por data editor; la segunda opción es usando las siguientes ventanas:

Ventana

file → import Excel → → localización del archivo (Browse) → señale el nombre del archivo a importar → open → activación de la opción: primera fila corresponde al nombre de las variables → okey.

1.2.2 Lectura de datos .tex

Cuando los datos están en archivos .tex, como por ejemplo, los que se encuentran en bloc de notas, se utilizan las siguientes ventanas:

Ventana

file → import → unformatted tex data → → localización del archivo (Browse) y digitación de las variables en su orden → okey.

1.2.3 Lectura de datos en formato .dta

Cuando los datos ya se encuentran en formato Stata, se pueden leer directamente utilizando las siguientes ventanas:

Ventana

file → open → → señale el nombre del archivo .dta → open.

1.3 Construcción de bases de datos

La construcción de bases de datos en Stata puede lograrse manualmente o a través de la combinación de bases ya existentes. En esta sección se explican ambas formas y sus variantes.

1.3.1 Introducción de datos de forma manual

Para introducir los datos manualmente a Stata se da clic en el menú data editor, el cual trae una hoja de cálculo similar a una hoja Excel, allí se digitán los datos para cada una de las variables que conforman la base de datos. Por defecto *default*, el Stata denomina las variables *var1*, *var2*, etc. Para colocar los nombres que hemos de asignar a cada una de las variables, el tipo de variable y, de ser el caso, sus *labels* para las categorías de las variables cualitativas, se recurre al último botón de la barra de herramientas: *variables manager*.

1.3.2 Combinación de bases de datos

En ocasiones el estudiante o investigador necesita combinar una base de datos inicial, que se encuentra en formato Stata .dta y que se denomina *master dataset*, con otra base de datos, también en formato Stata .dta, que ya había sido guardada, la cual se denomina *using dataset*. Un caso típico son las encuestas desarrolladas por el DANE, las cuales están organizadas por módulos, y cada uno de ellos contiene determinada información, por ejemplo: características del hogar, educación, empleo, salud, etc. En este caso es necesario combinar los módulos o archivos en uno solo para poder realizar los análisis estadísticos y económéticos de interés. Los módulos o archivos se pueden combinar de forma horizontal, *Join by joins* y vertical.

En forma horizontal

La combinación de base de datos horizontal adiciona a la base de datos inicial (*master dataset*), variables que se encuentran en otra base de datos (*using dataset*), esto se logra ordenando en forma ascendente una variable clave común a las dos bases y utilizando el comando *merge*, el cual permite adicionar variables a la base inicial (*master dataset*). Para adicionar más variables es necesario abrir el archivo original y seguir los pasos que se muestran a continuación:

Ventana

1. file → open → señalar el nombre del archivo .dta → abrir → open.
2. Ordene en forma ascendente la variable clave o común a los archivos a combinar: data → sort → seleccione la variable clave → .
3. Grabe el archivo con la variable clave ordenada: file → save.

Cierre las variables del archivo anterior con el comando clear y abra el archivo dos (*using dataset*) que contiene las variables a adicionar, repita el procedimiento realizado al archivo inicial (*master dataset*) y ciérrelo.

Teniendo los dos archivos con la variable clave ordenada se procede a combinarlos, se sugiere que sea al archivo original (*master dataset*) al que se le adicionan las variables, y seguir los pasos que se muestran a continuación:

Ventana

1. file → open → señale el nombre del archivo .dta → open.
2. Una vez abierto el archivo continúe con data → combine data set → merge two dataset.
3. Elegir el tipo de combinación *merge* que requiere su base de datos, por defecto escogemos (*one to one on key variables*).
4. Ubique la variable clave y abra la ruta donde se localiza la segunda base de datos (*using dataset*) → okay.

De esta manera queda un nuevo archivo con las variables iniciales más las variables adicionadas, como se observa en la [Figura 1.3](#).

La opción que trae el programa por default, (*one to one on key variables*) se conoce como **merge 1:1**, y se utiliza cuando el individuo se observa una sola vez en las bases de datos a combinar.

merge(1:1)

Base de datos 1

Individuo	X1	X2
1	27	0,4
2	23	0,3
3	24	0,6
4	29	0,45
5	30	0,37
6	19	0,7
7	25	0,39
8	19	0,5

Base de datos 1

Individuo	X3	X4
1	120	2550
2	115	1937
3	132	2054
4	127	2820
5	123	2143
6	141	1950
7	129	2800
8	115	2456

+

Base de datos nueva

Individuo	X1	X2	X3	X4
1	27	0,4	120	2550
2	23	0,3	115	1937
3	24	0,6	132	2054
4	29	0,45	127	2820
5	30	0,37	123	2143
6	19	0,7	141	1950
7	25	0,39	129	2800
8	19	0,5	115	2456

Figura 1.3: merge (1:1).

Siguiendo los pasos anteriores, se tienen otras opciones para combinar bases de datos en forma horizontal, como:

- **Merge 1:m:** en esta opción la variable clave identifica solo una observación en el *master dataset*, mientras que el *using dataset* identifica más de una.
- **Merge m:1:** en esta opción la variable clave identifica más de una observación en el *master dataset*, mientras que el *using dataset* identifica solo una.
- **Merge m:m:** en esta opción la variable clave identifica más de una observación tanto en el *master dataset* como en el *using dataset*

Para ilustrar el merge 1:m y m:1, supongamos que tenemos una base de datos con información individual por departamentos, en esta base de datos cada observación contiene información acerca de un solo departamento. Suponemos ahora que tenemos otra base de datos con información sobre todas las empresas del país clasificadas por departamento. Para poder unir ambas bases de datos tenemos las dos formas explicadas anteriormente: la primera, 1:m buscará de un departamento (uno) específico todas las empresas (muchas) que le correspondan al *using dataset*, siendo el *master dataset* los departamentos. La segunda forma, m:1, se usará cuando de todas las empresas (muchas) se busca a cuál departamento (uno) pertenece; muchas empresas buscarán un departamento en específico, siendo el *master dataset* las empresas y el *using dataset* los departamentos. La [Figura 1.4](#) muestra los resultados de merge 1:m y m:1.

merge(m:1)

Master			+	Using		=	Base de datos nueva			
Empresa	Departamento	Z		Departamento	X		Empresa	Departamento	Z	X
1	3	89		1	1254		1	3	89	895
2	4	101		2	2567		2	4	101	2115
3	1	295		3	895		3	1	295	1254
4	4	750		4	2115		4	4	750	2115
5	3	159					5	3	159	895
6	2	860					6	2	860	2567

merge(1:m)

Master		+	Using			=	Base de datos nueva			
Departamento	X		Empresa	Departamento	Z		Departamento	X	Empresa	Z
1	1254		1	3	89		3	895	1	89
2	2567		2	4	101		4	2115	2	101
3	895		3	1	295		1	1254	3	295
4	2115		4	4	750		4	2115	4	750
			5	3	159		3	895	5	159
			6	2	860		2	2567	6	860

Figura 1.4: Ejemplo de combinación de datos merge (m:1) (1:m)

Para ilustrar el m:m se hace referencia a la Encuesta de Calidad de Vida que realiza el DANE. La encuesta está dividida en módulos, los cuales contienen información de las familias y de los miembros que la componen. Si el investigador necesita integrar los módulos de vivienda, educación, empleo y salud, debe utilizar el comando merge m:m pues en los módulos se encuentran tanto los códigos de las familias como los integrantes con sus respectivas variables. La Figura 1.5 presenta un ejemplo básico de la combinación de datos merge m:m.

merge(m:m)

Master				Using				Base de datos nueva					
Id. Familia	Secuencia	X1	X2	Id. Familia	Secuencia	X3	X4	Id. Familia	Secuencia	X1	X2	X3	X4
1000	1	23	0,24	1000	1	20	0,79	1000	1	23	0,24	20	0,79
1000	2	28	0,29	1000	2	25	0,25	1000	2	28	0,29	25	0,25
1000	3	37	0,17	1000	3	34	0,2	1000	3	37	0,17	34	0,2
1001	1	35	0,73	1001	1	32	0,61	1001	1	35	0,73	32	0,61
1001	2	30	0,73	1001	2	27	0,86	1001	2	30	0,73	27	0,86
1002	1	38	0,45	1002	1	29	0,97	1002	1	38	0,45	29	0,97
1002	2	38	0,14	1002	2	34	0,84	1002	2	38	0,14	34	0,84
1002	3	33	0,13	1002	3	31	0,03	1002	3	33	0,13	31	0,03
1002	4	36	0,11	1002	4	33	0,92	1002	4	36	0,11	33	0,92
1003	1	28	0,1					1003	1	28	0,1	-	-
1003	2	25	0,84					1003	2	25	0,84	-	-

Figura 1.5: Ejemplo de combinación de datos merge (m:m).

En algunas bases de datos se requiere tener más de una variable clave, esto no debe generar confusión, pues lo único adicional es explicitar las variables claves a considerar, una seguida de la otra.

Join by joins

Puede considerarse como un caso particular del **merge**, pues adiciona las variables del archivo *using dataset* al *master dataset* solo cuando las observaciones aparecen en ambas bases. Al igual que en **merge**, las bases deben estar ordenadas de acuerdo a una o varias variables clave; para lograr esto se procede como se describió en la subsección Combinación de bases de datos.

Para ilustrar la combinación de datos Join by joins, suponemos que tenemos información acerca de familias y sus miembros, distribuida en dos bases de datos: la primera contiene la información de los padres de familia y sus características (*master dataset*), en la otra base de datos tenemos la información de los hijos y unas características (*using dataset*). Para unir ambas bases de datos se sugiere que sea al archivo original (*master dataset*) que se le adicionan las variables, siguiendo los pasos que se muestran a continuación:

Ventana

1. file → open → ↗
2. Señale el nombre del archivo .dta → open → data → combine data set → form all pairwise combinations within groups → ↗

Ubique la ruta donde se localiza el archivo (Browse) a unir y especifique la variable clave (*join observations*), en nuestro caso, X1 seguido de Okey. Es importante aclarar que la nueva base de datos solo colocará las familias que se crucen entre las bases de datos. Dado el caso que haya una familia que esté en una base de datos y en la otra no, esta será omitida, la [Figura 1.6](#) muestra el resultado de los pasos efectuados.

Joinby

Master			
Código Familia	Código padres	X1	X3
1087	115	80	7894
1087	116	65	2578
1385	122	59	1920
1385	123	68	8756
1554	156	91	3567
1554	157	83	9985

+

Using				
Código Familia	Código de hijo	X1	X2	
1087	879	76	7,5	
1087	878	56	6,4	
1385	945	65	3,9	
1554	991	60	8,8	
1554	992	53	7,9	

=

Base de datos nueva				
Código Familia	Código Padres	X1	X3	Código Hijos
1087	115	80	7894	879
1087	115	80	7894	878
1087	116	65	2578	879
1087	116	65	2578	878
1385	122	59	1920	945
1385	123	68	8756	945
1554	156	91	3567	991
1554	156	91	3567	992
1554	157	83	9985	991
1554	157	83	9985	992

Figura 1.6: Ejemplo de combinación de datos join by Joins

En forma vertical

En ocasiones se tienen dos archivos con diferentes unidades observadas a las cuales se les han medido las mismas características y se desea colocar todas las unidades observadas en una sola base de datos, en esta situación se utiliza el comando **append**. La sintaxis de este comando es mucho más sencilla, pues solo se tiene que nombrar la base de datos que se desea anexar. La ruta para adicionar observaciones a una base de datos inicial es:

Ventana

1. file → open → ↗
2. Ubicar la base a la cual se le quieren añadir observaciones, entonces hacer: open → data combine datasets → append datasets ↗
3. Ubique el archivo desde el cual va adicionar las nuevas observaciones al archivo inicial (Browse) → okey.

El resultado es presentado en la [Figura 1.7](#).

Append

Base de datos 1

Individuo	X1	X2
1	27	0,4
2	23	0,3
3	24	0,6
4	29	0,45
5	30	0,37
6	19	0,7
7	25	0,39
8	19	0,5

Base de datos 2

Individuo	X1	X2
9	23	0,35
10	25	0,4
11	29	0,54
12	29	0,7
13	28	0,39
14	31	0,6
15	27	0,5
16	23	0,4

Base de datos nueva

Individuo	X1	X2
1	27	0,4
2	23	0,3
3	24	0,6
4	29	0,45
5	30	0,37
6	19	0,7
7	25	0,39
8	19	0,5
9	23	0,35
10	25	0,4
11	29	0,54
12	29	0,7
13	28	0,39
14	31	0,6
15	27	0,5
16	23	0,4

Figura 1.7: Ejemplo de combinación de datos append.

Convertir datos de wide a long o viceversa

reshape convierte datos que se presentan de manera ancha *wide* a datos de manera larga *long* y viceversa. Para poder hacer uso del comando reshape y convertir datos, es necesario establecer si los datos están en forma *long* o forma *wide*, y determinar la observación lógica (i) y la subobservación (j). La ruta a seguir es:

Ventana

data → create or change data → Other variable-transformation commands → Convert data between wide and long.

Los comandos para convertir los datos a forma *wide* o *long* son:

Comando

(long → wide): reshape wide consumo, i(id) j(año)
 (wide → long): reshape long consumo, i(id) j(año)

La [Figura 1.8](#) muestra el resultado de la conversión de *wide* a *long* y viceversa.

Base de datos long

Individuo	Año	Raza	Consumo
1	2010	0	3400
1	2011	0	3550
1	2012	0	3700
2	2010	0	2950
2	2011	0	3100
2	2012	0	3200
3	2010	1	3350
3	2011	1	3370
3	2012	1	3900

Base de datos wide

Individuo	Sexo	Consumo(2010)	Consumo(2011)	Consumo(2012)
1	0	3400	3550	3700
2	0	2950	3100	3200
3	1	3350	3370	3900

Figura 1.8: Ejemplo de datos *long* y *wide*.

1.4 Análisis de datos

Para iniciar un análisis de datos es necesario haber cumplido con el proceso de depuración, excluyendo aquellas observaciones no válidas para el análisis.

Para conocer los datos, un comando muy útil es **describe**, el cual da una descripción de los datos, como el tipo, la etiqueta, el nombre de variable y el formato.

1.4.1 Obtención de base datos condensada

En ocasiones, la base de datos está conformada por unidades que a su vez agrupan otras unidades más pequeñas, por ejemplo, familias y miembros de la familia. Cuando deseamos calcular ciertas estadísticas de la familia, a partir de las características de los miembros de quienes se tiene información, se utiliza el comando **collapse**, esto permite convertir la base de datos en memoria en una base de datos de indicadores, como edad promedio de la familia, ingreso promedio de la familia, años promedio de estudio de la familia, etc. La ruta a seguir es:

Ventana

data → create or change data → Other variable-transformation commands → Make dataset of means, medians, etc.

NOTA: Es importante destacar que las variables deben ser exclusivamente numéricas. Para filtrar otro tipo de variables se coloca después de **collapse** la palabra **clist**.

La [Figura 1.9](#) muestra el resultado del proceso realizado anteriormente. Usando comandos se tiene:

Comando

`collapse (mean) edad educación Ingreso, by (familia)`

El comando anterior crea una base de datos con cuatro variables (hogar, edad, educación e ingreso) y una observación por hogar, la cual representa el promedio de cada variable por familia, como se observa en la [Figura 1.9](#).

Familia	Código padres	Edad	Educación	Ingreso (miles)
1025	132	42	18	3450
1026	145	32	16	2980
1025	133	39	15	1560
1027	147	50	21	4200
1026	146	37	18	2349
1027	149	53	18	1980

collapse
→

Familia	Edad promedio	Educación promedio	Ingreso promedio
1025	40,5	16,5	2505
1026	34,5	17	2664,5
1027	51,5	19,5	3090

Figura 1.9: Ejemplo de comando **collapse**.

NOTA: Si tenemos una base de datos en la que alguna variable tenga omisión de datos, al final del comando colocamos **cw**, lo que permitirá realizar el análisis estadístico solo con los datos existentes y omitirá aquellos en los que no haya datos.

1.4.2 Análisis descriptivo de datos

Stata contiene algunos comandos y opciones para realizar análisis estadístico y restringir la muestra a observaciones que cumplen con ciertas condiciones. Algunos de estos comandos son:

- **summarize**: muestra estadísticas descriptivas de las variables. En el caso de ser una variable cualitativa, el Stata excluye del análisis la categoría base de la variable cualitativa digitando, por ejemplo, **summarize i.sexo**. Si se requiere la información descriptiva para la categoría base, se debe digitar **summarize ibn.sexo**, así se mostrará la información de todas las categorías de la variable cualitativa.
- **summarize e***: muestra las estadísticas descriptivas de todas las variables que empiecen con la letra e.
- **summarize x1 x2 x3, separator(1)**: inserta una línea entre las estadísticas descriptivas de cada variable.

- **inspect**: este comando indica la cantidad de valores positivos, negativos, ceros y faltantes.
- **tabulate**: con este comando se obtienen las tablas de frecuencia.
- **codebook**: Stata muestra percentiles, desviación estándar, media, etc.

NOTA: El comando se coloca seguido de la variable o variables a estudiar, y luego el condicional que se desea utilizar, por ejemplo: **summarize experiencia if experiencia > 5, detail**. El comando anterior resumirá la información de todos los individuos que tengan una experiencia superior a 5 años. Otra opción es realizar el análisis estadístico por grupos a través del comando **by**, por ejemplo: **summarize experiencia by (familia)**.

Por ventana, la mayoría de las estadísticas de los datos se obtienen de la siguiente forma:

Ventana

data → describe data → summary statistics clic Seleccione las variables okey.

1.5 Creación y modificación de variables en la base de datos

Construida la base de datos, en ocasiones se hace necesario crear nuevas variables a partir de las ya existentes en la base; en este caso se hace uso del comando **generate** acompañado de la expresión matemática correspondiente. Por ejemplo, la función **minceriana** incorpora como variable explicatoria la experiencia al cuadrado. Si en la base original se tiene la experiencia pero no su cuadrado, se debe generar esta variable de la siguiente manera:

Comando

generate experiencia2=experiencia*experiencia

Por ventana se realiza de la siguiente manera:

Ventana

1. data → create or change → create new variable → 
2. Nombrar la nueva variable y explicitar la expresión matemática correspondiente.

A continuación se presentan algunos comandos que permiten eliminar, mantener, generar, recodificar o renombrar variables de la base de datos.

- **drop**: elimina las variables seleccionadas.
- **keep**: mantiene las variables seleccionadas.
- **egen**: este comando es una extensión del comando **generate**. Cuando se desea crear una variable con estructuras complejas, en ocasiones es muy difícil y hasta imposible hacerlo con el comando **generate**, apareciendo así el comando **egen**, el cual permite generar variables con estructuras complejas. Por ejemplo, **egen becas= mean(calificaciones) if calificaciones > 4,5**. El comando anterior creará una nueva variable llamada **becas** que calculará la nota media de los alumnos, pero solo para aquellos que tengan notas superiores a 4,5.
- **recode**: es una extensión de **replace** que recodifica, entre otras, a variables cuantitativas en grupos o clases. Por ejemplo, si una variable cuantitativa puede tomar valores de 1 a 30, es posible agruparla en categorías, así: las observaciones que estén entre 1 y 10 tomarán el valor de 1, las observaciones que estén entre 11 y 20 tomarán el valor de 2 y así sucesivamente. **recode riesgo (1/10=1) (11/19=2) (20/30=3)**, **generate (nriesgo)**, el comando anterior recodificará la variable denominada **riesgo** en una nueva variable llamada **nriesgo**, la cual tomará valores de 1, 2 y 3.

Por otro lado, en algunas bases de datos los valores **missing** se identifican con 99 o 999, dado que para el Stata los valores **missing** se representan por punto (.). Dichos valores (99 o 999) deben recodificarse para poder ser tratados por Stata como **missing** a través del siguiente comando: **recode X (99=.)**, de esta forma, donde Stata localice un 99 en la variable X lo reemplazará por un punto (.) y lo tratará, por consiguiente, como un valor **missing**. Observe que la recodificación en este caso se hace sobre la misma variable, mientras en el primer ejemplo la recodificación se hace sobre una nueva variable.

- **rename**: renombrar una variable de la base de datos. Ejemplo: **rename price precio**. Para este caso, la variable original price se renombra con la palabra precio.
- **generate trend = _n**: genera una variable de tendencia denominada trend

NOTA: Los comandos **drop** y **keep** permiten condicionales a las variables que se quieren modificar eliminar o mantener en la base de datos. Por ejemplo, **drop edu*** borra todas las variables que empiezan por edu; **drop if salario>4000** borra todas las observaciones que tengan salario mayor de 4000; **keep in 5/15** considera para el análisis desde la observación 5 hasta la 15; **keep if salario>3500 y dum==1** mantiene aquellas observaciones en las cuales el salario es >3500 y dum=1, eliminando el resto de observaciones.

1.6 Apéndices

Apéndice 1. merge 1:1.

Archivo do

```
*Borre todas las bases de datos existentes
clear
* Importe el archivo base de datos 1 xls, ordene la variable
* clave en forma ascendente y guárdelo en formato.dta
import excel "J:\Stata y econometría\Base de datos 1.xlsx"
sheet("Hoja1") firstrow sort individuo
save "J:\Stata y econometría\base1.dta"
file J:\Stata y econometría\base1.dta saved
*cierre el archivo anterior e importe el archivo base de datos 2. XLS
*ordene la variable clave en forma ascendente y guárdelo
*en formato.dta
clear
import excel "J:\Stata y econometría\Base de datos 2.xlsx"
sheet("Hoja1") firstrow
sort individuo
save "J:\Stata y econometría\base2.dta"
file J:\Stata y econometría\base2.dta saved
*cierre el archivo base2 y abra el archivo base1a.dta
clear
use "J:\Stata y econometría\base1.dta", clear
*Combine el archivo base1 con el archivo base 2
merge 1:1 individuo using "J:\Stata y econometría\base2.dta"
```

Apéndice 2. merge 1:m.

Archivo do

```
*importe su archivo master, donde la primera fila es para
*nombre de las variables
import excel "J:\Stata y econometría\master 1:m.xlsx"
sheet("Hoja1") firstrow
*Ordene en orden ascendente su variable clave
sort Departamento
*Grabe y cierre el archivo
save "J:\Stata y econometría\master1:m.dta"
clear
*Importe su archivo Using, donde la primera fila para nombre
*de las variables
import excel "J:\Stata y econometría\Using 1:m.xlsx"
sheet("Hoja1") firstrow
*Ordene en orden ascendente su variable clave
sort Departamento
*Grabe y cierre el archivo
save "J:\Stata y econometría\Using 1:m.dta"
```

```
clear  
*Abra el archivo master  
use "J:\Stata y econometria\master1:m.dta"  
clear  
*Combine archivo .dta master 1:m con archivo using 1:m
```

Modelo de regresión lineal múltiple con datos transversales

Temas Tratados

2.1 Estimación por mínimos cuadrados ordinarios (MCO) para datos transversales

- 2.1.1 Explicación de la salida de un modelo de regresión lineal múltiple estimado en Stata
- 2.1.2 Interpretación de los coeficientes de regresión estimados
- 2.1.3 Inferencia en el modelo de regresión lineal múltiple
- 2.1.4 Hipótesis sobre una combinación lineal o conjunto de combinaciones lineales de los coeficientes de regresión
- 2.1.5 Formas funcionales de un modelo

En econometría se trabaja con tres tipos de datos: corte transversal, series de tiempo y longitudinales o de panel. Los datos de corte transversal se pueden entender como aquellos que se recogen en un momento específico del tiempo; un ejemplo típico son las encuestas para estudios de mercado. Un modelo que utiliza este tipo de datos emplea el subíndice i en su estructura. En esta sección se presenta el modelo de regresión lineal múltiple con datos de corte transversal bajo el supuesto de muestreo aleatorio, como lo sugiere (Wooldridge, 2010), dejando para el [capítulo 5](#) los modelos que utilizan series de tiempo, y para el 8 los modelos de datos de panel.

El análisis de regresión múltiple es una de las herramientas utilizadas por la econometría para estudiar la dependencia de una variable llamada dependiente de otra u otras denominadas independientes o explicatorias. El objeto del análisis de regresión, en ocasiones, es predecir el valor de la variable dependiente, dados ciertos valores a las variables independientes.

2.1 Estimación por mínimos cuadrados ordinarios (MCO) para datos transversales

Caso de estudio. Supóngase que se deben identificar cuáles son los factores que determinan o explican el ingreso por salario mensual de los individuos. La herramienta más apropiada para resolver este problema es el modelo de regresión lineal múltiple, en el que se considere como variable dependiente el ingreso por salario en función de variables, como los años de estudio, la experiencia y la experiencia al cuadrado. La experiencia tiene un efecto positivo sobre el salario hasta cierto punto, a partir del cual el ingreso por salario empezará a disminuir. El efecto positivo de la experiencia sobre el salario es cada vez menor, lo que refleja un crecimiento marginal decreciente de este con respecto a la experiencia. De acuerdo con lo anterior, el modelo de regresión lineal múltiple planteado será:

$$\text{ingresosalario}_i = \beta_1 + \beta_2(\text{añosdeestudio})_i + \beta_3\text{experiencia}_i + \beta_4\text{experiencia}_i^2 + U_i$$

Para estimar por MCO, en el modelo anterior se utilizan datos de la encuesta del mercado laboral en Colombia del periodo 1967-1970, elaborada por la universidad de los Andes.

Stata, para estimar una regresión por MCO, utiliza el comando **regress** seguido por la variable dependiente y el conjunto de variables independientes o explicatorias, en nuestro caso se usa el siguiente comando:

Comando

```
regress Ingresosalario añosdeestudio exper exper2
```

Antes de estimar el modelo anterior, se debe generar una nueva variable que represente la variable experiencia al cuadrado (exper^2), para esto se hace uso del comando **generate** presentado en la [subsección 1.5](#), de la siguiente manera:

Comando

```
generate exper2=exper * exper
```

El comando **regress** presenta una serie de posibilidades al realizar la estimación, las cuales se describen a continuación:

1. **Estimar la regresión sin generar el cuadrado de la variable exper.**

Comando

```
regress Ingresosalario añosdeestudio exper c.exper c.#c.exper
```

2. Hacer la regresión por cada categoría de sexo.

Comando

```
by sexo: regress IngresoSalario añosdeestudio exper exper2
```

Este comando hará la regresión por cada categoría de sexo, es decir, realizará una estimación para las mujeres y otra estimación para los hombres. La variable sexo tiene que estar previamente ordenada de manera ascendente (*sort*).

3. **Estimar la regresión para todas las observaciones donde una variable sea superior a un valor.** Este comando estimará la regresión para todas las observaciones donde la variable añosdeestudio sea superior a 10.

Comando

```
regress Ingresosalario anosdeestudios exper exper2, if añosdeestudio > 10
```

4. **Agrupar las variables explicatorias del modelo con global \$xlist.** En nuestro caso *añosdeestudio*, *exper*, *exper2*, de tal manera que al estimar el modelo se reemplaza el conjunto de variables explicatorias por el comando \$xlist, es decir, para obtener la estimación se procede de la siguiente manera:

Comando

```
regress Ingresosalario $xlist
```

5. **Estimar el modelo de regresión sin constante.**

Comando

```
regress Ingresosalario añosdeestudio exper exper2, noconstant
```

6. **Construir el intervalo de confianza al nivel deseado, por defecto está al 95%.**

Comando

```
regress Ingresosalario añosdeestudio exper exper2, Level(#)
```

7. **Eliminar el encabezado de la salida.**

Comando

```
regress IngresoSalario añosdeestudio exper exper2, noheader
```

8. **Reportar los coeficientes beta estandarizados.**

Comando

```
regress Ingresosalario añosdeestudio exper exper2, beta
```

El modelo estimado y la explicación detallada de la información retornada por el Stata se muestra en la [Figura 2.1](#).

. reg Ingresosalario Añosdeestudio exper exper2

Source	SS	df	MS	Number of obs =	23035
Model	2.1689e+10	3	7.2297e+09	F(3, 23031) =	87.90
Residual	1.8927e+12	23031	82179073.4	Prob > F =	0.0000
Total	1.9144e+12	23034	83109988.9	R-squared =	0.0113
				Adj R-squared =	0.0112
				Root MSE =	9065.3

Ingresosalario	Coeff.	Std. Err.	t	P> t	[95% Conf. Interval]
Añosdeestudio	254.8772	15.7467	16.19	0.000	224.0125 285.7417
exper	80.70024	15.75757	5.12	0.000	49.81435 111.5861
exper2	-1.107897	.287891	-3.85	0.000	-1.672103 -.5436111
_cons	-649.5444	227.3081	-2.86	0.004	-1095.084 -204.0053

Figura 2.1: Estimación MCO en Stata.

2.1.1 Explicación de la salida de un modelo de regresión lineal múltiple estimado en Stata

Encabezado

la regresión estimada en la cual figura la variable dependiente ingresosalario seguida por las variables independientes añosdeestudio, exper y exper2.

En la parte superior izquierda

- **Model-SS= 2.1689e+10:** representa la suma explicada de cuadrados o **SEC**.
- **Residual-SS= 1.8927e+12:** representa la suma residual de cuadrados o **SRC**.
- **Total-SS= 1.9144e+12:** representa la suma total de cuadrados o **STC**.
- **Model-df= 3:** representa los grados de libertad asociados a la suma explicada de cuadrados o **SEC**.
- **Residual-df= 23031:** representa los grados de libertad asociados a la suma residual de cuadrados o **SRC**.
- **Total-df= 23034:** representa los grados de libertad asociados a la suma total de cuadrados o **STC**.
- **Model-MS= 7.2297e+09:** representa el cuadrado medio de la regresión, y resulta de dividir la suma de cuadrados del modelo entre sus grados de libertad.
- **Residual-MS= 82179073.4:** representa el cuadrado medio residual, y resulta de dividir la suma residual de cuadrados entre sus grados de libertad. Es importante destacar que este cuadrado medio es la estimación de la varianza del error.
- **Total-MS= 83109988.9:** representa el cuadrado medio del total, y resulta de dividir la suma total de cuadrados entre sus grados de libertad.

En la parte superior derecha

- **number of obs=23035:** representa el número de observaciones utilizadas en la estimación.
- **F(3.23031)=87.98:** representa el estadístico F para significancia global, el cual se define como:

$$F((k-1), (n-k)) = \frac{SEC/k - 1}{SRC/n - k} \quad (2.1)$$

siendo k el número de coeficientes de regresión en el modelo y n el número de observaciones utilizadas en la estimación, en nuestro ejemplo, $k=4$ y $n=23035$; por lo tanto, el número de grados de libertad para la F es 3 grados en el numerador y 23031 en el denominador.

- **Prob > F=0.0000:** es el valor p asociado al estadístico F .
- **R-squared=0.0113:** representa el coeficiente de determinación o el R^2 , definido como:

$$R^2 = \frac{SEC}{STC} \quad (2.2)$$

utilizado para medir el ajuste del modelo a los datos.

- **Adj-R-squared= 0.0112:** representa el coeficiente de determinación ajustado, definido como:

$$\overline{R^2} = 1 - (1 - R^2) \frac{n - 1}{n - k} \quad (2.3)$$

y utilizado para comparar dos modelos que tienen la misma variable dependiente, pero diferente número de variables explicativas.

- **Root-RSE= 9065.3:** representa la desviación estándar estimada del error, es decir, la raíz cuadrada de la varianza estimada del error.

En la parte inferior, el Stata utiliza el método de mínimos cuadrados para realizar la estimación del modelo, este método consiste en hallar los estimadores de los coeficientes de regresión, β_1 , β_2 , β_3 y β_4 , de tal manera que la suma de los cuadrados de los residuos sea mínima, es decir, $\sum_{i=0}^n \hat{u}_i^2$ mínima.

En la primera columna se encuentra el nombre de las variables explicativas y el intercepto representado por **cons**; en la segunda columna, los valores numéricos de las estimaciones mínimo cuadráticas de cada uno de los coeficientes de regresión asociados a cada variable explicatoria, así como el valor numérico estimado del intercepto **cons**; por lo tanto, la curva de regresión estimada será:

$$\begin{aligned} \text{ingresosalario}_i &= -649,54 + 254,87 \times \text{añosdeestudio}_i + 80,70 \times \text{exper}_i - 1,107 \times \text{exper2}_i \\ \text{ee} &= (227,30) & (15,74) & (15,75) & (0,28) \\ \text{Valor p} &= (0,00) & (0,00) & (0,00) & (0,00) \\ t &= (-2,85) & (16,18) & (5,12) & (-3,84) \end{aligned}$$

2.1.2 Interpretación de los coeficientes de regresión estimados

- $\hat{\beta}_1$ = Ante un aumento de un año de estudio, se espera que en promedio el ingreso por salario mensual aumente 254.87 pesos. *ceteris paribus*.
- $\hat{\beta}_2$ = Se espera que en promedio el ingreso por salario mensual en el primer año de experiencia se incremente 80.70 pesos. *ceteris paribus*.
- $\hat{\beta}_3$ = Dado que esta depende del nivel de experiencia, este coeficiente se utiliza frecuentemente para encontrar el punto de quiebre.
- R^2 = El 1,13% de la variación total del ingreso por salario es explicada por el modelo.

2.1.3 Inferencia en el modelo de regresión lineal múltiple

Dado que las estimaciones obtenidas corresponden a una muestra seleccionada aleatoriamente, la pregunta es: ¿Cómo se generalizan los resultados a toda la población? El siguiente paso, después de la estimación puntual de los coeficientes de regresión poblacionales por MCO, es la generalización de los resultados a toda la población, es decir, la inferencia estadística.

Dado que los estimadores $\hat{\beta}_i$ de los coeficientes de regresión poblacionales β_i son variables aleatorias, pues sus valores se modifican para

cada muestra seleccionada aleatoriamente, para poder adelantar inferencias (estimación por intervalos e hipótesis) alrededor de los coeficientes de regresión poblacionales, se necesita conocer la distribución de probabilidad de dichos estimadores, la cual se obtiene a partir del conocimiento de la distribución de probabilidad del término de error o perturbación. Tradicionalmente y partiendo del teorema del límite central, se asume que el término de error o perturbación sigue una distribución normal, por tanto, los estimadores seguirán esa misma distribución con una media $E(\hat{\beta}_i) = \beta_i$ y una varianza $\sigma_{\hat{\beta}_i}^2$, la cual es una medida de la precisión del estimador.

Para probar si las perturbaciones o el error siguen una distribución normal en el ejemplo anterior, el Stata tiene tres pruebas incorporadas: **Skewness and kurtosis normality test**, **Shapiro-Wilk normality test** y **Shapiro Wilk-Francia normality test**. Para utilizar la prueba **Shapiro Wilk** se recomienda que el número de observaciones sea inferior a 5000, y para un número mayor de observaciones se usa **Skewness and kurtosis normality test**, los comandos asociados a las pruebas anteriores son **sktest**, **swilk** y **sfrancia**, respectivamente.

La decisión es, si el valor p asociado al estadístico **sktest**, **swilk** > a no rechaza **H₀**, los errores siguen una distribución normal, en caso contrario, se rechaza. En este caso, las inferencias deben tomarse con cuidado, pues la inferencia exacta basada en los estadísticos t y F requiere normalidad. No obstante, actualmente se demuestra que cuando el tamaño de muestra es grande, los estimadores MCO satisfacen la normalidad asintótica. Como en nuestro caso el número de observaciones es 23035, se asume normalidad asintótica en el error y, por tanto, en los estimadores MCO.

Para realizar la prueba **sktest**, **swilk** o **sfrancia** se deben seguir los pasos que se indican en las ventanas:

swilk

Ventana

1. Correr el modelo siguiendo: Statistics → Postestimation → predictions → residuals ↗
2. Asigne un nombre a los residuos del modelo → ok.
3. Statistics → Summaries → Tables and tests → Distributionals plots and tests → (escoger uno de los siguientes testings) Skewness and kurtosis normality test, Shapiro-Wilk normality test o Shapiro-Francia normality test.

Los comandos para la estimación de estos tests después de la estimación del modelo son los siguientes:

Comando

1. predict residuos, residuals
2. sktest residuos
3. swilk residuos
4. sfrancia residuos

Significancia individual

Con relación a la inferencia estadística alrededor de los coeficientes de regresión poblacionales, el Stata incorpora en la salida mostrada en la [Figura 2.1](#), lo que se conoce como pruebas de significancia individual, cuya estructura es:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

La tercera columna de la tabla inferior de la [Figura 2.1](#) presenta el error estándar $S_{\hat{\beta}_i}$ (Std. Err.) o la raíz cuadrada de la varianza estimada de $\hat{\beta}_i$ asociada a cada coeficiente de regresión estimado. Así mismo, en la cuarta columna se presenta la razón t, la cual se calcula como $t = \frac{\hat{\beta}_i - 0}{S_{\hat{\beta}_i}}$, después de fijar los puntos críticos, si el t calculado que aparece en la tabla cae en la región de rechazo, la hipótesis nula de no significancia individual se rechaza, es decir, el β_i poblacional es significativo, lo que implica que la variable asociada al mismo contribuye a la explicación de la variable dependiente.

Una forma alterna de realizar las pruebas de significancia individual es a través del valor p (*p value*), que es el valor exacto de cometer el error tipo I o el nivel de significancia más bajo al cual puede rechazarse una hipótesis nula. El criterio utilizado es: si el valor p es menor que el nivel de significancia α , se rechaza la hipótesis nula de no significancia individual, mientras si el valor p es mayor al nivel de significancia α , no se rechaza. La quinta columna de la tabla inferior en la [Figura 2.1](#) muestra que todos los coeficientes de regresión del modelo planteado son significativamente diferentes de 0, pues el valor p en todos los casos es menor que un nivel de significancia del 5%, por tanto, las variables explicatorias contribuyen a explicar individualmente la variable dependiente.

Finalmente, Stata en la sexta columna de la tabla inferior en la [Figura 2.1](#) presenta un intervalo de confianza del 95% asociado a cada coeficiente de regresión poblacional. Estos intervalos también pueden ser usados para probar las hipótesis de significancia individual, pues si dicho intervalo contiene el valor 0, no se rechaza la hipótesis nula de no significancia, mientras que se rechaza en caso contrario, el nivel de significancia en esta prueba es del 5%.

En el caso analizado del modelo ingreso por salario, ninguno de los intervalos construidos por default por el Stata contiene el 0, es decir, se rechaza la hipótesis nula de no significancia individual para todos los coeficientes de regresión del modelo, lo cual coincide con el resultado obtenido usando el valor p.

Prueba de hipótesis de significancia global

Ante la pregunta: ¿Cómo se prueba la significancia global del modelo? La respuesta es: las hipótesis clásicas de significancia individual, como su nombre lo indica, permiten probar la significancia de cada coeficiente de regresión individualmente. Para probar la hipótesis de significancia global o conjunta de los coeficientes de regresión se plantean las siguientes hipótesis:

$$H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$$
$$H_1: Al \text{ } menos \text{ } un \beta_i \neq 0, \quad i \leq k$$

Para esta prueba se utiliza el estadístico F, el cual puede ser definido con la [ecuación \(2.4\)](#) o [\(2.5\)](#).

$$F = \frac{SEC/k - 1}{SRC/n - k} \tag{2.4}$$

$$F = \frac{R^2/K - 1}{(1 - R^2)/n - K} \tag{2.5}$$

Donde

- K= número de parámetros, y
- n= número de observaciones.

En la salida del Stata, para el caso analizado del salario por ingreso, se observa tanto el valor del estadístico F calculado = 87.98 como su valor p asociado=0.000, esto implica que se rechaza la hipótesis nula de no significancia global, es decir, al menos uno de los coeficientes de regresión poblacionales β_i que acompañan a las variables explicatorias es diferente de 0, o sea que el modelo es significativo globalmente. Recuerde que en la hipótesis nula de esta prueba no se considera el coeficiente de regresión asociado al intercepto β_1 .

2.1.4 Hipótesis sobre una combinación lineal o conjunto de combinaciones lineales de los coeficientes de regresión

Un interrogante que surge en esta subsección es: ¿Cómo se prueban hipótesis sobre combinaciones lineales entre los coeficientes de regresión poblacionales del modelo, como por ejemplo, $H_0: \beta_2 = \beta_3$? Estas hipótesis se pueden probar de dos maneras: la primera, llevando la igualdad a la forma de una combinación lineal, es decir, $H_0: \beta_2 - \beta_3 = 0$, y luego utilizar el estadístico t para una combinación lineal. En el caso considerado, dicho estadístico t viene dado por la [ecuación \(2.6\)](#).

$$t = \frac{(\widehat{\beta}_2 - \widehat{\beta}_3) - (\beta_2 - \beta_3)}{S_{\widehat{\beta}_2 - \widehat{\beta}_3}} \tag{2.6}$$

Donde:

$$S_{\widehat{\beta}_2 - \widehat{\beta}_3} = \sqrt{V(\widehat{\beta}_2 - \widehat{\beta}_3)} = \sqrt{V(\widehat{\beta}_2 + \widehat{\beta}_3) - 2Cov(\widehat{\beta}_2, \widehat{\beta}_3)}$$

Este estadístico se calcula a partir de las estimaciones de los coeficientes de regresión y de la matriz estimada de varianzas y covarianzas del vector de coeficientes de regresión estimados, con la cual se obtiene el error estándar de la combinación lineal planteada. Para obtener la matriz estimada de varianzas y covarianzas, después de haber estimado el modelo, se utiliza el comando **vce**. La [Figura 2.2](#) muestra los resultados para la combinación lineal planteada en el caso de estudio.

. vce

Covariance matrix of coefficients of regress model

e (V)	Añosdeestu-o	exper	exper2	_cons
Añosdeestu-o	247.95862			
exper	61.698022	248.30089		
exper2	-.70988399	-4.2941023	.08288125	
_cons	-2312.5516	-2945.528	44.094399	51668.98

Figura 2.2: Resultado para la combinación lineal.

Una segunda forma de probar la hipótesis de interés anterior es a través de mínimos cuadrados restringidos enfoque F, donde la hipótesis nula será:

$$H_0 : \beta_2 = \beta_3$$

El estadístico utilizado por este enfoque es un F, el cual viene dado la [ecuación \(2.7\)](#)

$$F = \frac{(SCR_0 - SCR_a)/m}{SRC_a/(n - k)} \quad (2.7)$$

Donde:

- SCR_0 = suma de cuadrados residuales bajo la hipótesis nula o restringida.
- SCR_a = suma de cuadrados residuales bajo la hipótesis alterna o no restringida.
- m = número de restricciones o el número de coeficientes de regresión que desaparecen del modelo original al imponer la restricción.

Para obtener el valor calculado del estadístico F calculado, Stata procede de la siguiente manera: primero, se estima el modelo común y corriente, sin ningún tipo de restricción, y se obtiene SCR_a ; segundo, se impone la restricción planteada en la hipótesis nula y se estima el modelo restringido para obtener SCR_0 ; tercero, se determina m como el número de restricciones o el número de coeficientes de regresión que se pierden al restringir el modelo; finalmente, se determina n (número de observaciones consideradas en la estimación) y k (número de parámetros en el modelo original).

El programa Stata permite probar la hipótesis $H_0 : \beta_2 = \beta_3$, a través del comando **test**, el cual hace uso de mínimos cuadrados restringidos enfoque F, de la siguiente forma:

Comando

```
test añosdeestudio=exper
```

En otras ocasiones, la teoría económica sugiere la existencia de un conjunto de relaciones entre los coeficientes de regresión poblacionales, las cuales deben ser contrastadas, como por ejemplo: $H_0 : \beta_1 = \beta_2$ y $H_0 : \beta_4 = 0$.

El estadístico de prueba para las hipótesis anteriores viene dado por la [ecuación \(2.8\)](#).

$$F = \frac{(R\hat{\beta} - r)^t [R(X^t X)^{-1} R^t]^{-1} (R\hat{\beta} - r)/m}{\hat{u}^t \hat{u}/T - k} \quad (2.8)$$

La [ecuación \(2.8\)](#) es equivalente a la [ecuación \(2.9\)](#):

$$F = (\widehat{R\beta} - r)^t [\widehat{\sigma}_u^2 R(X^t X)^{-1} R^t]^{-1} (\widehat{R\beta} - r) / m \quad (2.9)$$

Donde se tiene:

$$[R] = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} [r] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} [\widehat{\beta}] = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \\ \widehat{\beta}_4 \end{bmatrix}$$

Podemos utilizar el comando **test** de Stata para probar las dos hipótesis planteadas como sigue:

Comando

```
test (añosdeestudio=exper) (exper2)
```

La [Figura 2.3](#) presenta los resultados de las dos hipótesis planteadas.

```
. test (Añosdeestudio= exper) (exper2)

( 1) Añosdeestudio - exper = 0
( 2) exper2 = 0

F( 2, 23031) = 120.60
Prob > F = 0.0000
```

Figura 2.3: Prueba de hipótesis múltiple.

2.1.5 Formas funcionales de un modelo

En esta subsección surge la siguiente pregunta: ¿Cómo se calcula a través de un modelo econométrico una elasticidad o una tasa de crecimiento? La respuesta es: en los cursos de Econometría básica, y en ocasiones en avanzada, se acostumbra a trabajar con modelos lineales en los parámetros, independientemente de si hay o no una relación lineal entre las variables. Existen algunos modelos de interés para los econometristas, a partir de los cuales es posible calcular, por ejemplo, la elasticidad del precio de la demanda de un bien o la tasa de crecimiento de una variable económica. A continuación se presentan las formas funcionales de un modelo lineal más frecuentemente utilizadas

- **El modelo lineal en su forma funcional log-log**

En este modelo la forma funcional que relaciona a las variables Y y X viene dada por la [ecuación \(2.10\)](#):

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i} \quad (2.10)$$

Aplicando logaritmo natural a ambos lados se tiene que:

$$\ln(Y_i) = \ln \beta_1 + \beta_2 \ln X_i + u_i \quad (2.11)$$

Reescribiendo la expresión anterior se obtiene:

$$\ln(Y_i) = \alpha + \beta_2 \ln X_i + u_i \quad (2.12)$$

Lo que implica que el modelo es lineal en α y β , y lineal en los logaritmos de las variables. En estos modelos, los coeficientes de regresión asociados a las variables explicatorias, en este caso β_2 , representan una elasticidad, es decir, cambios porcentuales en Y ante cambios porcentuales en X.

■ El modelo lineal en su forma funcional log-lin

En este modelo la forma funcional que relaciona a las variables Y y X viene dada por la [ecuación \(2.13\)](#):

$$Y_i = e^{\beta_1 + \beta_2 X_i + u_i} \quad (2.13)$$

Aplicando logaritmo natural a ambos lados se obtiene:

$$\ln(Y_i) = \beta_1 + \beta_2 X_i + u_i \quad (2.14)$$

Lo que implica que el modelo es lineal en β_1 y β_2 . En estos modelos, los coeficientes de regresión asociados a las variables explicatorias, en este caso β_2 , representan cambios porcentuales en Y ante cambios absolutos en X. Cuando X es una variable de tendencia, el coeficiente asociado representa una tasa de crecimiento. Debe tenerse en cuenta que para la interpretación las pendientes deben multiplicarse por 100.

A continuación se presenta la aplicación e interpretación de los dos modelos anteriores, tomando como referencia los datos correspondientes del modelo ingreso por salario estimado anteriormente. Es importante aclarar que antes de realizar las estimaciones se deben generar los logaritmos de las variables originales utilizando el comando **generate**, así:

Comando

```
generate Insalario= log(Ingresosalario)
```

Comando

```
generate Inañosdeestudio= log(añosdeestudio)
```

Comando

```
generate Inexperiencia= log(exper)
```

Comando

```
generate exper2= exper * exper
```

$$1. \text{ Insalario}_i = \beta_1 + \beta_2 \ln \text{añosdeestudio}_i + \beta_3 \ln \text{Inexperiencia}_i + u_i$$

$$2. \text{ Insalario}_i = \beta_1 + \beta_2 \ln \text{añosdeestudio}_i + \beta_3 \ln \text{experiencia}_i + \beta_4 \ln \text{exper}^2 + u_i$$

Interpretaciones

$$1. \text{ Insalario}_i = 3,978 + 0,8823 \ln \text{añosdeestudio}_i + 0,3939 \ln \text{Inexperiencia}_i$$

- β_1 : Ante un aumento de 1% en años de estudio se espera que en promedio el salario aumente aproximadamente 0,88%, *ceteris paribus*.
- β_3 : Ante un aumento de 1% en la experiencia se espera que en promedio el salario aumente aproximadamente 0,39%, *ceteris paribus*.

Debe quedar claro que las dos interpretaciones anteriores no tienen mucho sentido.

$$2. \text{ Insalario}_i = 4,73 + 0,152 \text{ añosdeestudio}_i + 0,067 \text{ exper}_i - 0,0009 \text{ exper}^2_i$$

- β_1 : Ante un aumento de un año de estudio se espera que en promedio el salario aumente aproximadamente 15,2%, *ceteris paribus*.

- β_1 : El primer año de experiencia incrementa el salario aproximadamente 6,7%, *ceteris paribus*.

A diferencia de las interpretaciones en el modelo log-log, las dos anteriores tienen todo un sentido económico y permiten contribuir en el diseño de política económica.

Análisis de regresión con información cualitativa

Temas Tratados

- 3.1 Una variable cualitativa como explicatoria**
 - 3.1.1 Caso I. La variable cualitativa modifica solo el intercepto
 - 3.1.2 Caso II. Cuando la dummy modifica solo la pendiente de años de educación
 - 3.1.3 Caso III. Cuando la dummy modifica el intercepto y la pendiente
- 3.2 Más de una variable cualitativa como explicatoria**
 - 3.2.1 Caso I. Más de una variable cualitativa sin interacción
 - 3.2.2 Caso II. Más de una variable cualitativa como explicatoria con interacción

Este capítulo se inicia planteando el siguiente interrogante: ¿Es posible incorporar e interpretar variables cualitativas como el sexo, la raza y la profesión en un modelo de regresión lineal múltiple? La respuesta es que sí es posible incorporar dichas variables, haciendo uso de las denominadas variables dummy. Para introducir este tema, se parte del hecho de que adicional a la educación y la experiencia, el sexo también puede ser considerado como un determinante del ingreso por salario, de acuerdo con estudios como los presentados por (De Fanelli, 1989) y (Peñas, 2002) sobre discriminación salarial.

3.1 Una variable cualitativa como explicatoria

Una variable cualitativa puede tomar más de una categoría, como por ejemplo, el estado civil; sin embargo, en este manual se considera el caso cuando dicha variable cualitativa toma solo dos categorías, como el sexo. Cuando se tienen en cuenta más de dos categorías el análisis se aborda de manera similar.

3.1.1 Caso I. La variable cualitativa modifica solo el intercepto

En este tipo de modelos, la variable cualitativa explicatoria ingresa al modelo en forma aditiva y su coeficiente asociado mide el diferencial en intercepto frente a la categoría base o aquella que recibe el valor de 0.

Por ejemplo, para el modelo de regresión de los determinantes del ingreso por salario en su forma funcional log-lin, se ha incorporado, como variable explicatoria, el sexo. El modelo queda expresado en la [ecuación \(3.1\)](#).

$$\text{Insalario}_i = \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 + \alpha \cdot \text{sexo}_i + u_i \quad (3.1)$$

Recuerde, como se dijo al inicio del [Capítulo 2](#), que la experiencia al cuadrado se justifica porque la experiencia tiene un efecto positivo, en este caso, sobre el Insalario hasta cierto punto, a partir del cual el Insalario empezará a disminuir. El efecto positivo de la experiencia sobre el Insalario es cada vez menor, lo que refleja un crecimiento marginal decreciente de este con respecto a la experiencia.

Ahora bien, a partir del modelo anterior se sugiere generar dos funciones de regresión para el Insalario, una para las mujeres cuando sexo=0, [ecuación \(3.2\)](#), y otra para los hombres cuando sexo=1, [ecuación \(3.3\)](#).

Función salario mujer

$$\begin{aligned} E(\text{Insalario} | \text{sexo} = 0, \text{añosdeestudio}_i, \text{exper}_i) &= \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i \\ &\quad + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \end{aligned} \quad (3.2)$$

Función salario hombre

$$\begin{aligned} E(\text{Insalario} | \text{sexo} = 1, \text{añosdeestudio}_i, \text{exper}_i) &= \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i \\ &\quad + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 + \alpha \end{aligned} \quad (3.3)$$

$$E(\text{Insalario}|\text{sexo} = 1, \text{añosdeestudio}_i, \text{exper}_i) = (\beta_1 + \alpha) + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \quad (3.4)$$

Como se observa, la función Insalario hombres difiere de la función Insalario mujeres solo por el término α , el cual representa el diferencial entre ambas funciones. Obsérvese que las pendientes de las variables añosdeeducación, exper y exper al cuadrado, medidas por β_2 , β_3 y β_4 , son las mismas tanto para hombres como para mujeres, es decir, el efecto del sexo sobre Insalario se deja sentir solo en el intercepto, siempre y cuando este diferencial sea significativo.

La [Figura 3.1](#) muestra que dado un nivel de años de estudio y experiencia, el salario promedio de los hombres es superior al de las mujeres en 52% (una estimación más precisa es de 68.2%). Las pendientes se interpretan de manera similar, como se hizo anteriormente en los modelos lineales, cuya forma funcional es log-lin.

. reg Insalario Añosdeestudio exper exper2 Sexo

Source	SS	df	MS	Number of obs	=	23035
Model	9359.23317	4	2339.80829	F(4, 23030)	=	2529.48
Residual	21303.0736	23030	.925014053	Prob > F	=	0.0000
Total	30662.3068	23034	1.33117595	R-squared	=	0.3052
				Adj R-squared	=	0.3051
				Root MSE	=	.96178

Insalario	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Añosdeestudio	.1390696	.0017047	81.58	0.000	.1357283 .1424109
exper	.0550404	.0017023	32.33	0.000	.0517038 .0583771
exper2	-.0007709	.0000308	-25.05	0.000	-.0008313 -.0007106
Sexo	.522935	.0135981	38.46	0.000	.4962818 .5495882
_cons	4.652406	.0242138	192.14	0.000	4.604946 4.699867

Figura 3.1: Dummy cambio de intercepto.

3.1.2 Caso II. Cuando la dummy modifica solo la pendiente de años de educación

Si se asume que la incorporación de la variable cualitativa sexo afecta solamente la pendiente de añosdeestudio, la formulación del modelo se muestra en la [ecuación \(3.5\)](#).

$$\begin{aligned} \text{Insalario}_i = & \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \\ & + \alpha \cdot \text{sexo} * \text{añosdeestudio}_i + u_i \end{aligned} \quad (3.5)$$

Igual que en el caso I, se generan dos funciones de regresión para el Insalario, una para las mujeres cuando sexo=0, como se observa en la [ecuación \(3.6\)](#), y otra para los hombres cuando sexo=1, expresada en la [ecuación \(3.8\)](#).

Función salario mujer

$$\begin{aligned} E(\text{Insalario mensual}|\text{sexo} = 0, \text{añosdeestudio}_i, \text{exper}_i) = & \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i \\ & + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \end{aligned} \quad (3.6)$$

Función salario hombre

$$E(\ln\text{salario} | \text{sexo} = 1, \text{añosdeestudio}_i, \text{exper}_i) = \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 + \lambda \cdot \text{añosdeestudio} \quad (3.7)$$

Sacando factor común añosdeestudio y agrupando términos, se obtiene:

$$E(\ln\text{salario} | \text{sexo} = 1, \text{añosdeestudio}_i, \text{exper}_i) = \beta_1 + (\beta_2 + \lambda) \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \quad (3.8)$$

Como se observa, la función salario hombres ([ecuación \(3.8\)](#)) difiere de la función salario mujeres ([ecuación \(3.6\)](#)) solo en la pendiente de la variable añosdeestudio por el término λ , el cual representa el diferencial entre ambas funciones por cada año adicional de estudio, siempre y cuando este diferencial sea significativo. Observe que el intercepto en ambas funciones es el mismo. Para la estimación del modelo con cambio solamente en la pendiente, se genera una nueva variable denominada **sexestu**, la cual se obtiene multiplicando las variables **añosdeestudio** por el sexo. Generada dicha variable se estima el modelo común y corriente con el comando:

Comando

```
regress lnSalario añosdeestudio exper exper2 sexestu
```

La [Figura 3.2](#) muestra que para un nivel dado de experiencia, un año de estudio adicional de los hombres incrementa su salario 3.74% más que lo que le incrementa ese mismo año adicional de estudio a una mujer. Esta diferencia resulta ser significativa.

```
. reg lnSalario Añosdeestudio exper exper2 sexestu
```

Source	SS	df	MS	Number of obs	=	23035
Model	8335.044	4	2083.761	F(4, 23030)	=	2149.35
Residual	22327.2628	23030	.96948601	Prob > F	=	0.0000
Total	30662.3068	23034	1.33117595	R-squared	=	0.2718
				Adj R-squared	=	0.2717
				Root MSE	=	.98462
<hr/>						
lnSalario	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Añosdeestudio	.1227862	.0023128	53.09	0.000	.1182531	.1273194
exper	.0602122	.0017534	34.34	0.000	.0567755	.0636489
exper2	-.0008211	.0000317	-25.91	0.000	-.0008832	-.000759
sexestu	.037494	.001991	18.83	0.000	.0335915	.0413966
_cons	4.85227	.0254501	190.66	0.000	4.802386	4.902154

Figura 3.2: Estimación del modelo con cambio en la pendiente.

3.1.3 Caso III. Cuando la dummy modifica el intercepto y la pendiente

Si se asume que la incorporación de la variable cualitativa sexo afecta tanto al intercepto como a la pendiente de los años de estudio, la [ecuación \(3.9\)](#) representa la formulación del modelo.

$$\ln\text{salario}_i = \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 + \alpha_1 \cdot \text{sexo}_i + \alpha_2 \cdot (\text{sexo} * \text{añosdeestudio})_i + u_i \quad (3.9)$$

Igual que en los casos I y II, se generan dos funciones de regresión para el Insalario, una para las mujeres cuando sexo=0, [ecuación \(3.10\)](#), y otro para los hombres cuando sexo=1, [ecuación \(3.12\)](#).

Función salario mujer

$$E(\ln\text{salario}|\text{sexo} = 0, \text{añosdeestudio}_i, \text{exper}_i) = \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \quad (3.10)$$

Función salario hombre

$$E(\ln\text{salario}|\text{sexo} = 1, \text{añosdeestudio}_i, \text{exper}_i) = \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 + \alpha_1 + \alpha_2 \cdot \text{añosdeestudio}_i \quad (3.11)$$

Sacando factor común años de estudio y agrupando términos, se obtiene:

$$E(\ln\text{salario}|\text{sexo} = 1, \text{añosdeestudio}_i, \text{exper}_i) = (\beta_1 + \alpha_1) + (\beta_2 + \alpha_2) \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \quad (3.12)$$

Como se observa, la función salario hombres difiere de la función salario mujeres tanto en el intercepto en una cantidad α_1 , como en la pendiente de añosdeestudio en una cantidad α_2 , siempre y cuando estos diferenciales sean significativos.

La Figura 3.3 muestra que para un determinado nivel de experiencia, un año de estudio adicional de los hombres incrementa su salario 6.45%, por debajo del incremento que representa ese mismo año adicional de educación para las mujeres. Por otro lado, para un nivel de experiencia dado, cuando los años de educación son 0, los hombres tienen un salario promedio 88% superior al de las mujeres (una estimación más precisa es de 141%). Obsérvese que los diferenciales en intercepto y pendiente son significativos.

. reg lnSalario Añosdeestudio exper exper2 Sexo sexestu						
Source	SS	df	MS	Number of obs = 23035		
Model	9717.5054	5	1943.50108	F(5, 23029) = 2136.90		
Residual	20944.8014	23029	.909496783	Prob > F = 0.0000		
Total	30662.3068	23034	1.33117595	R-squared = 0.3169		
				Adj R-squared = 0.3168		
				Root MSE = .95368		
lnSalario	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Añosdeestudio	.1804604	.0026844	67.22	0.000	.1751987	.1857221
exper	.0588087	.0016986	34.62	0.000	.0554793	.0621381
exper2	-.0008357	.0000307	-27.22	0.000	-.0008958	-.0007755
Sexo	.8860065	.0227254	38.99	0.000	.8414633	.9305498
sexestu	-.0645081	.0032502	-19.85	0.000	-.0708787	-.0581375
_cons	4.394266	.0273063	160.93	0.000	4.340744	4.447788

Figura 3.3: Estimación del modelo con cambio en intercepto y pendiente.

3.2 Más de una variable cualitativa como explicatoria

Si bien a veces se requiere incorporar más de una variable cualitativa como explicatoria, en este manual dichas variables asumen solo dos valores. A continuación se estudiarán los casos, cuando existe y no existe interacción entre dichas variables.

3.2.1 Caso I. Más de una variable cualitativa sin interacción

El modelo formulado incorpora, en ocasiones, más de una variable cualitativa como explicatoria en forma aditiva. Si se asume que no existe

interacción entre dichas variables, el efecto de estas se dejará sentir solo en el intercepto. El modelo para Insalario, considerado en este capítulo, se modifica agregando dos variables cualitativas como explicatorias, el sexo y la zona, donde el sexo toma el valor de 0 cuando es mujer y 1 cuando es hombre, y la zona toma el valor de 0 si no es Bogotá y 1 si es Bogotá. El modelo se describe en la [ecuación \(3.13\)](#).

$$\begin{aligned} \text{Insalario} = & \beta_1 + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \\ & + \alpha_1 \cdot \text{sexo}_i + \alpha_2 \cdot \text{zona}_i + u_i \end{aligned} \quad (3.13)$$

Para el modelo anterior se consideraron, para ilustración, solo dos funciones de regresión para el Insalario, una cuando sexo=0 y zona=0, [ecuación \(3.14\)](#), y otra cuando sexo=1 y zona=1, [ecuación \(3.15\)](#).

Función salario mujer y no Bogotá

$$\begin{aligned} E(\text{Insalario} | \text{sexo} = 0, \text{zona} = 0, \text{añosdeestudio}_i, \text{exper}_i) = & \beta_1 \\ & + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \end{aligned} \quad (3.14)$$

Función salario hombre y Bogotá

$$\begin{aligned} E(\text{Insalario} | \text{sexo} = 1, \text{zona} = 1, \text{añosdeestudio}_i, \text{exper}_i) = & (\beta_1 + \alpha_1 + \alpha_2) \\ & + \beta_2 \cdot \text{añosdeestudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 \end{aligned} \quad (3.15)$$

La [Figura 3.4](#) muestra que para un determinado nivel de experiencia y años de estudio, los hombres ganan 52.82% más que las mujeres (la estimación más precisa es de 68.2%), independientemente de la ciudad de residencia, en este caso, si viven o no en Bogotá.

. reg Insalario Añosdeestudio exper exper2 Sexo Zona						
Source	SS	df	MS	Number of obs = 23035		
Model	9429.58295	5	1885.91659	F(5, 23029) = 2045.46		
Residual	21232.7239	23029	.921999386	Prob > F = 0.0000		
Total	30662.3068	23034	1.33117595	R-squared = 0.3075		
				Adj R-squared = 0.3074		
				Root MSE = .96021		
Insalario	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Añosdeestudio	.1381984	.0017048	81.06	0.000	.1348569	.14154
exper	.0549545	.0016996	32.33	0.000	.0516233	.0582858
exper2	-.0007689	.0000307	-25.02	0.000	-.0008292	-.0007087
Sexo	.5282151	.0135894	38.87	0.000	.501579	.5548512
Zona	.1373314	.0157219	8.74	0.000	.1065155	.1681473
_cons	4.54569	.027086	167.82	0.000	4.4926	4.598781

Figura 3.4: Dummys sin efecto interacción.

De la misma manera, una persona que resida en Bogotá tiene un salario superior en 13.7% (siendo una estimación más precisa 14.68%), independientemente de si es hombre o mujer. En resumen, un hombre que resida en Bogotá gana 65.7% (una estimación más precisa es de 92.99%) más que una mujer que no resida en Bogotá.

3.2.2 Caso II. Más de una variable cualitativa como explicatoria con interacción

La formulación de este modelo se expresa en la [ecuación \(3.16\)](#), tomamos en cuenta la multiplicación de las variables sexo y zona, lo cual implica que hay interacción entre ellas, es decir, el salario de un hombre o una mujer podría ser mayor o menor dependiendo de la zona.

$$\text{lnsalario mensual} = \beta_0 + \beta_1 \cdot \text{añosdeestudio}_i + \beta_2 \cdot \text{exper}_i + \beta_3 \cdot \text{exper}_i^2 + \alpha_1 \cdot \text{sexoi} \\ + \alpha_2 \cdot \text{zona}_i + \gamma(\text{sexoi} * \text{zona}_i) + u_i \quad (3.16)$$

La Figura 3.5 muestra para un determinado nivel de experiencia y años de estudio, que los hombres que viven fuera de Bogotá ganan 61% (una estimación más precisa es de 84%), más que las mujeres que residen fuera de Bogotá. Por otro lado, una mujer residente en Bogotá gana 20.3% (una estimación más precisa es de 22.5%), más que una mujer que vive fuera de Bogotá. Finalmente, un hombre que reside en Bogotá gana 71.2% (una estimación más precisa es de 103.8%), más que una mujer que reside fuera de Bogotá. Debe tenerse en cuenta que estos diferenciales son significativos.

. reg lnsalario Añosdeestudio exper exper2 Sexo Zona zonasex

Source	SS	df	MS	Number of obs	=	23035
Model	9438.41443	6	1573.06907	F(6, 23028)	=	1706.79
Residual	21223.8924	23028	.921655914	Prob > F	=	0.0000
Total	30662.3068	23034	1.33117595	R-squared	=	0.3078
				Adj R-squared	=	0.3076
				Root MSE	=	.96003

lnsalario	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Añosdeestudio	.138225	.0017045	81.09	0.000	.134884 .141566
exper	.0548747	.0016995	32.29	0.000	.0515437 .0582058
exper2	-.0007677	.0000307	-24.98	0.000	-.000828 -.0007075
Sexo	.6102288	.0297751	20.49	0.000	.5518676 .6685899
Zona	.2035734	.0265522	7.67	0.000	.1515293 .2556175
zonasex	-.1018677	.0329082	-3.10	0.002	-.16637 -.0373654
_cons	4.492394	.0320908	139.99	0.000	4.429493 4.555294

Figura 3.5: Dummys con efecto interacción.

NOTA: Cuando en la interpretación se menciona una estimación más precisa, se está haciendo uso del ajuste que debe realizarse a los coeficientes de las variables dummy en los modelos lineales en la forma funcional log-lin. Dicho ajuste consiste en tomar el antilogaritmo del coeficiente de la dummy restarle 1 y a dicha diferencia multiplicarla por 100.

Problemas econométricos

Temas Tratados

- 4.1 Multicolinealidad**
 - [4.1.1 Multicolinealidad perfecta](#)
 - [4.1.2 Alta multicolinealidad](#)
 - [4.1.3 Pruebas para detectar alta multicolinealidad](#)
 - [4.1.4 Medidas remediales a la alta multicolinealidad](#)
- 4.2 Heterocedasticidad**
 - [4.2.1 ¿Cómo detectar heterocedasticidad?](#)
 - [4.2.2 ¿Cómo solucionar heterocedasticidad?](#)
- 4.3 Especificación**
 - [4.3.1 Pruebas para variables omitidas y forma funcional incorrecta](#)
 - [4.3.2 Pruebas para selección de modelos](#)
 - [4.3.3 Criterios de selección de modelos](#)

A continuación se presentan tres problemas econométricos que surgen con frecuencia en el trabajo aplicado cuando se emplean diversos tipos de datos, entre ellos los de corte transversal.

4.1 Multicolinealidad

Las variables económicas que se incorporan como explicatorias en un modelo de regresión, por lo general provienen de un sistema económico dado, en consecuencia, es natural encontrar que dichas variables estén correlacionadas y, aunque no quitan las propiedades a los estimadores MELI (bajo la premisa de que se cumplen los supuestos del modelo), sí generan inconvenientes desde el punto de vista de la aplicación, dependiendo del grado de dicha correlación. De acuerdo con lo anterior, el problema no es que exista o no correlación entre las variables explicatorias, sino del grado o fuerza de esa correlación. En el trabajo hay tres casos: multicolinealidad perfecta o exacta, alta multicolinealidad y baja multicolinealidad.

4.1.1 Multicolinealidad perfecta

La multicolinealidad perfecta se presenta cuando una o varias variables explicatorias pueden expresarse como combinación lineal exacta de las demás, en estos casos los coeficientes de regresión no se pueden estimar individualmente, lo que se puede estimar son combinaciones lineales de dichos coeficientes; además, las varianzas de los estimadores son infinitas. La razón por la cual los coeficientes de regresión no pueden ser estimados individualmente es porque el estimador mínimo cuadrático del vector de coeficientes de regresión viene dado por la [ecuación \(4.1\)](#).

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad (4.1)$$

$$(X^t X)^{-1} = \frac{\text{Adjunta}(X^t X)}{[X^t X]} \quad (4.2)$$

Sabemos que $(X^t X)^{-1} = \text{Adjunta}(X^t X)/[X^t X]$, si existe multicolinealidad perfecta, entonces el determinante de $[X^t X] = 0$, implicando con ello que $\hat{\beta}$ y su varianza $\hat{\sigma}_u^2(\hat{\beta}) = \hat{\sigma}_u^2(X^t X)^{-1}$, no puedan ser estimados. Hay dos maneras de solucionar el problema: (i) eliminando una de las variables explicatorias que genera el problema, y (ii) hallando la relación lineal exacta que existe entre las variables explicatorias e introducirla en el modelo que presenta el problema de multicolinealidad perfecta. Esta última solución permite conocer las combinaciones lineales que pueden ser estimadas. Afortunadamente, la multicolinealidad perfecta no se presenta en la práctica, a no ser que sea un problema inducido por la manipulación de los datos.

4.1.2 Alta multicolinealidad

La alta multicolinealidad se presenta cuando ninguna variable explicatoria se puede expresar como combinación lineal exacta de las demás, pero entre ellas existe alta correlación, por lo general superior al 80%. En estos casos, los coeficientes de regresión pueden ser estimados individualmente, pero con errores estándar altos, lo cual genera que la mayoría de las razones t no rechacen la hipótesis nula de no significancia individual.

4.1.3 Pruebas para detectar alta multicolinealidad

Existen diferentes pruebas para detectar alta multicolinealidad, una de ellas es el factor de inflación de varianza centrado (VIF), que se define como:

$$VIF = \frac{1}{1 - R_j^2}$$

Donde R_j^2 es el coeficiente de determinación de la regresión entre X_j y el resto de las variables explicatorias en el modelo, un VIF mayor de 10 indica problemas de alta multicolinealidad, no obstante, en esta regla, un VIF alto no es condición necesaria ni suficiente para obtener errores estándar altos, pues, como se sabe, la varianza de un coeficiente de regresión estimado depende de tres factores: la varianza del error, la variabilidad del regresor correspondiente y el coeficiente de determinación entre el regresor X_j y las demás variables explicatorias.

La Figura 4.1 muestra los resultados obtenidos para el VIF centrado para el modelo ingreso por salario; como era de esperarse, el VIF asociado a la variable experiencia y experiencia al cuadrado es superior a 10. No obstante, dado que la multicolinealidad se define como la existencia de relaciones lineales entre las variables explicatorias, se puede decir que en el modelo de ingresos por salarios no existe multicolinealidad porque la relación entre experiencia y experiencia al cuadrado es no lineal. Sin embargo, si la relación entre experiencia y experiencia al cuadrado es muy alta, desde el punto de vista práctico, se dificulta realizar la estimación más precisa de los coeficientes de regresión.

La prueba de multicolinealidad en Stata por medio de ventanas se realiza de la siguiente manera, después de estimar el modelo:

Ventana

1. statistics → postestimation → reports and statistics → ↘
2. Busque en la caja: variance inflation factors for the independent variables(vif) → ok.

Por comando sería:

Comando

```
estat vif
```

```
. estat vif
```

Variable	VIF	1/VIF
exper	10.68	0.093657
exper2	10.27	0.097381
Añosdeestu~o	1.14	0.878500
Mean VIF	7.36	

Figura 4.1: Estimación del factor de inflación de varianza.

4.1.4 Medidas remediales a la alta multicolinealidad

Para algunos autores, la presencia de alta multicolinealidad no es problema, pues no se rompen las propiedades MELI de los MCO, asumiendo que los otros supuestos se cumplen; no obstante, el problema es más de orden práctico y aplicado, ya que la alta multicolinealidad, como ya se dijo, puede generar que la gran mayoría de los coeficientes de regresión resulten no significativos. Existen varias maneras de solucionar el problema (Gujarati y Porter, 2010), entre otras se tienen:

- Eliminar las variables altamente correlacionadas, teniendo cuidado de no romper los marcos teóricos en los cuales se fundamenta el modelo.
- Informacidos a priori que permita hallar la relación lineal que posiblemente existe entre los coeficientes de regresión, de tal manera que al introducir dicha relación en el modelo se elimine la alta multicolinealidad.
- Aumentar el tamaño de muestra para alcanzar una mayor variabilidad en las variables explicatorias, y de esta forma disminuir la varianza asociada al estimador correspondiente. En este punto se debe resaltar que los datos que se adicionen a los existentes deben haber sido recogidos siguiendo la misma metodología de los originales.

4.2 Heterocedasticidad

El modelo supone que la $\text{Var}(u_i) = \sigma^2$ para toda subpoblación i , es decir, es homocedástica. Sin embargo, en estudios que utilizan información de corte transversal, por lo general dicho supuesto se rompe, siendo $\text{Var}(u_i) = \sigma_i^2$, lo que implica que los estimadores MCO dejen de ser MELI en la medida que ya no tienen la varianza mínima dentro de los estimadores lineales e insesgados, no obstante, siguen siendo insesgados y consistentes.

4.2.1 ¿Cómo detectar heterocedasticidad?

Para detectar presencia de heterocedasticidad existen diferentes métodos, entre otros se tienen:

Método gráfico

Para detectar la heterocedasticidad por el método gráfico se debe proceder de la siguiente manera:

1. Estime el modelo por MCO.
2. Genere la variable \hat{y} siguiendo la ruta después de estimar el modelo:

Ventana

1. statistics → postestimation → predictions, residuals → 
2. Asigne el nombre a la predicción en el cuadro nueva variable, por ejemplo yhat → ok.

3. Genere la variable \hat{u} siguiendo la ruta después de estimar el modelo:

Ventana

1. statistics → postestimatio → predictions, residuals → 
2. active residuals y asigne el nombre a los residuos en el cuadro nueva variable, por ejemplo uhat → ok.

4. Genere el cuadrado de la variable uhat, o sea $uhat^2$.
5. Construya el gráfico siguiendo la ruta:

Ventana

Graphics → twoway graph → create → scatter uhat² yhat → ok.

Usando la línea de comando:

Comando

```
predict yhat,xb
```

Comando

```
predict uhat,residual
```

Comando

```
generate uhat2 = uhat * uhat
```

Comando

```
scatter uhat2 yhat
```

Si existe un patrón sistemático de comportamiento se puede sospechar que hay heterocedasticidad, en caso contrario, no se presentarán indicios de heterocedasticidad.

Otra forma de identificar gráficamente la presencia de heterocedasticidad es realizando el mismo gráfico anterior, pero esta vez colocando los cuadrados de los residuos contra cada una de las variables explicatorias, esto implica que se tendrán tantas gráficas como variables explicatorias se tengan en el modelo. Una ventaja de graficar los cuadrados de los residuos contra cada una de las variables explicatorias, es que permite identificar no solo cuál variable es la que está generando el problema de heterocedasticidad, sino la forma funcional con que dicha variable se relaciona con el error, lo que permite encontrar la transformación adecuada para eliminar el problema de heterocedasticidad.

Las [Figuras 4.2](#) y [4.3](#) no presentan un comportamiento completamente aleatorio, lo que sugiere presencia de heterocedasticidad en los errores. Dado que el método gráfico es un poco subjetivo, el Stata tiene un conjunto de tests para probar estadísticamente si existen o no problemas de heterocedasticidad. Dos de los tests más ampliamente utilizados son el test de Breusch-Pagan y el test de White, que se explicarán a continuación.

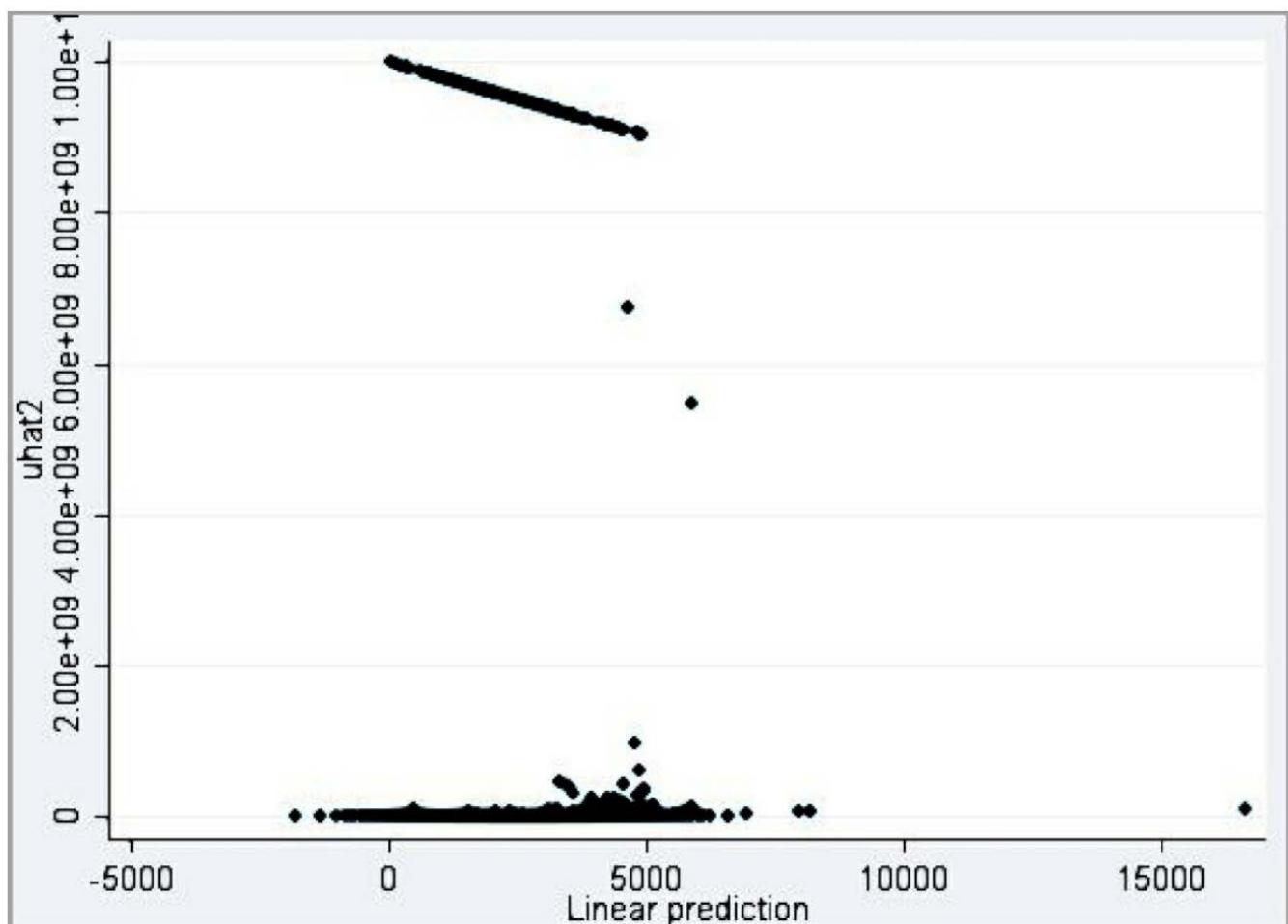


Figura 4.2: Gráfico de $uhat^2$ y $yhat$.

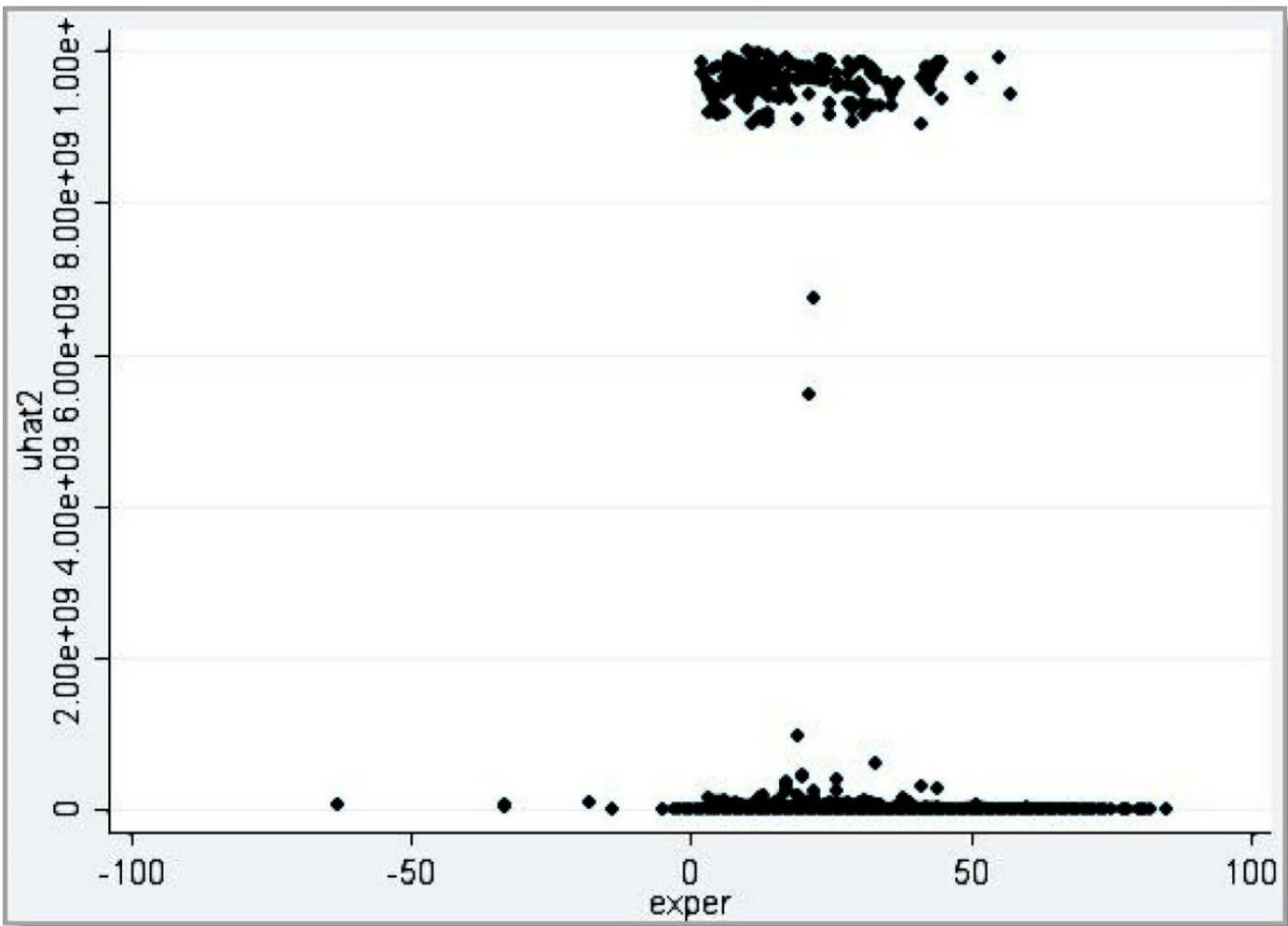


Figura 4.3: Gráfico cuadrado de residuos versus experiencia.

El test de Breusch-Pagan-Godfrey

Es un test del multiplicador de Lagrange, en el que la hipótesis nula es que no hay heterocedasticidad versus la hipótesis alterna, que considera que hay heterocedasticidad. La prueba en Stata después de estimar el modelo por medio de ventanas se realiza siguiendo esta ruta:

Ventana

Statistics → Postestimation → Reports and statistics → → test for heteroskedasticity → ok.

O por comando después de haber estimado el modelo:

Comando

`estat hettest`

```
. estat hettest  
  
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of Ingresosalario  
  
chi2 (1) = 412.47  
Prob > chi2 = 0.0000
```

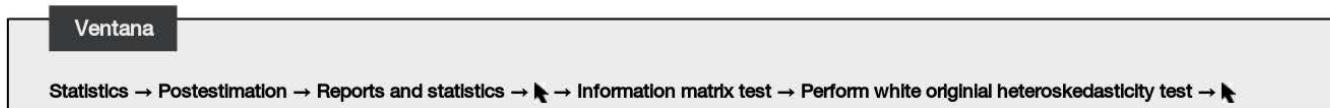
Figura 4.4: Prueba de Breusch-Pagan-Godfrey.

Los resultados presentados en la [Figura 4.4](#) muestran un valor $p=0.000$, lo que implica que los errores del modelo ingreso por salario presentan heterocedasticidad, pues se rechaza la hipótesis nula a un nivel de significancia del 5%.

El test de White

Es un test general para probar heterocedasticidad, donde la hipótesis nula es que no hay heterocedasticidad versus la hipótesis alterna, que considera que hay heterocedasticidad. La prueba consiste en realizar una regresión de los cuadrados de los residuos en función de las variables explicatorias del modelo, de sus cuadrados y de los productos cruzados entre dichas variables explicatorias. El Stata no muestra explícitamente esta regresión auxiliar.

La prueba en Stata después de estimar el modelo se realiza de la siguiente manera:



O por comando después de haber estimado el modelo:



Los resultados presentados en la [Figura 4.5](#) muestran un valor $p = 0,1126$, lo que implica que los errores del modelo ingreso por salario no presentan heterocedasticidad, pues no se rechaza la hipótesis nula a un nivel de significancia del 5%.

```
. estat imtest, white
```

White's test for Ho: homoskedasticity
against Ha: unrestricted heteroskedasticity

chi2(8) = **12.98**
Prob > chi2 = **0.1126**

Cameron & Trivedi's decomposition of LM-test

Source	chi2	df	p
Heteroskedasticity	12.98	8	0.1126
Skewness	208.05	3	0.0000
Kurtosis	-3.75e+09	1	1.0000
Total	-3.75e+09	12	1.0000

Figura 4.5: Test de White.

4.2.2 ¿Cómo solucionar heterocedasticidad?

Si existe evidencia de heterocedasticidad, se debe estimar el modelo por medio de mínimos cuadrados generalizados, es decir, aplicar MCO al modelo transformado adecuadamente de acuerdo con la forma funcional en que se relaciona la varianza del error y las variables explicatorias. En caso de que no se conozca dicha relación funcional, se deben utilizar estimadores robustos de White.

Para realizar la estimación con estimadores robustos de White se siguen los mismos pasos para estimar una regresión normal, pero se adiciona la palabra **robust**, anteponiendo una coma para separar las variables explicatorias del comando **robust**. El resultado se muestra en la Figura 4.6.

Una medida remedial utilizada con frecuencia en el trabajo empírico es la transformación logarítmica, la cual consiste en estimar el modelo que presenta heterocedasticidad tomando sus variables en logaritmo. Esta medida remedial elimina con frecuencia la heterocedasticidad porque la transformación logarítmica comprime las escalas en las cuales se miden las variables, reduciendo una diferencia entre dos valores de diez veces a una diferencia de dos veces. Una restricción que presenta la transformación logarítmica, es que no puede ser aplicada cuando uno o varios valores de las variables incorporadas en el modelo toman el valor de cero o negativo. Una salida a esta situación es sumando una cantidad K tanto a X como a Y, de tal manera que convierta los valores negativos o cero de las variables a cantidades positivas.

```
. reg Ingresosalario Aosdeestudio exper exper2, robust
```

Linear regression	Number of obs =	23035
	F(3, 23031) =	60.56
	Prob > F =	0.0000
	R-squared =	0.0113
	Root MSE =	9065.3

Ingresosal~o	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Aosdeestudio	254.8772	19.22764	13.26	0.000	217.1897	292.5646
exper	80.70024	14.24898	5.66	0.000	52.77129	108.6292
exper2	-1.107897	.2402431	-4.61	0.000	-1.578789	-.6370043
_cons	-649.5444	232.4307	-2.79	0.005	-1105.124	-193.9646

Figura 4.6: Estimadores robustos de White.

4.3 Especificación

Uno de los supuestos al diseñar un modelo econométrico es que está bien especificado, sin embargo, en el proceso de modelado es posible que dicho supuesto no se cumpla, pues se pueden cometer errores en la especificación como: omisión de variables relevantes, inclusión de variables irrelevantes, mala forma funcional, entre otros. Lo anterior tiene implicaciones en las propiedades de los estimadores MCO.

Como se presenta en los cursos de Econometría, la omisión de variables relevantes produce que los estimadores MCO sean sesgados e inconsistentes, además que las pruebas de hipótesis tradicionales no sean válidas. Por otro lado, la inclusión de variables irrelevantes genera que los estimadores MCO sean inefficientes, aunque son insesgados y consistentes y los procedimientos tradicionales de pruebas de hipótesis sean válidos. No obstante, no es recomendable incluir un sinnúmero de variables como explicatorias en el modelo, sino guiarse en la teoría subyacente, en la intuición y en un análisis de datos cuidadoso para obtener el modelo adecuado.

4.3.1 Pruebas para variables omitidas y forma funcional incorrecta

El Stata incorpora la prueba RESET, por que es una prueba general para detectar errores de especificación, como omisión de variables relevantes o mala forma funcional, lo que genera que los estimadores MCO sean sesgados e inconsistentes y que se invalide los procedimientos clásicos de la inferencia estadística. Esta prueba está fundamentada en una regresión aumentada con las potencias \hat{y} . Stata incorpora por defecto `default` las potencias de orden 2, 3 y 4. La hipótesis nula en esta prueba es que el modelo está bien especificado, es decir, que los coeficientes asociados a los términos adicionados son simultáneamente iguales a 0, mientras que la hipótesis alterna es que el modelo está mal especificado, es decir, que al menos uno de los coeficientes asociados a las variables adicionadas es diferente de 0. Para probar la hipótesis anterior se utilizan mínimos cuadrados restringidos enfoque F.

Para realizar la prueba RESET se estima primero el modelo y se sigue esta ruta:

Ventana

Statistics → Postestimation → Reports and statistics → ↗ → Ramsey regression specification.

Por comando después de haber estimado el modelo:

Comando

`estat ovtest`

La [Figura 4.7](#) muestra la aplicación de la prueba RESET. Los resultados rechazan la hipótesis nula, pues el valor p es menor que un nivel de

significancia del 5%. Una desventaja con RESET es que no proporciona ninguna indicación sobre cómo proceder si se rechaza el modelo, para autores como (Wooldridge, 2010), la prueba RESET es solo una prueba para detectar si la forma funcional del modelo planteado es adecuada o no.

```
. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of Ingresosalario
Ho: model has no omitted variables
F(3, 23028) =      22.50
Prob > F =      0.0000
```

Figura 4.7: Prueba de especificación.

4.3.2 Pruebas para selección de modelos

El investigador, en ocasiones se ve enfrentado a tener que seleccionar entre dos o más modelos; en esta situación se cuenta con un conjunto de pruebas por que pueden aplicarse dependiendo de la estructura de los modelos rivales, es decir, si son o no anidados.

Modelos anidados

Dos modelos son anidados cuando uno de ellos está contenido en el otro; para nuestro caso, supongamos que se debe seleccionar entre el modelo A y el modelo B.

$$A: \ln(\text{salario mensual}) = \beta_0 + \beta_1 \cdot \text{Años de estudio}_i + \beta_2 \cdot \text{exper}_i + \beta_3 \cdot \text{exper}_i^2 + \alpha_1 \cdot \text{sexo}_i \\ + \alpha_2 \cdot \text{zona}_i + \gamma \cdot (\text{sexo}^* \text{zona})_i + u_i$$

$$B: \ln(\text{salario mensual}) = \beta_1 + \beta_2 \cdot \text{Años de estudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 + \alpha \cdot \text{sexo}_i + u_i$$

Dado que el modelo B está contenido en el modelo A, es decir, son anidados, para la selección se utilizan mínimos cuadrados restringidos enfoque F. Las hipótesis son las siguientes:

$$H_0: \alpha_2 = \gamma = 0 \\ H_1: \text{Al menos uno } \neq 0$$

El estadístico de prueba viene dado por

$$F = \frac{SCR_{H_0} - SCR_{H_a}/M}{SCR_{H_a}/n - k} \quad (4.3)$$

Donde:

- SCR_{H_0} es la suma de cuadrados residuales bajo la hipótesis nula;
- SCR_{H_a} es la suma de cuadrados residuales bajo la hipótesis alterna;
- M es el número de restricciones;
- n el tamaño de la muestra;
- K el número de coeficientes de regresión en el modelo no restringido.

Con base en los resultados obtenidos anteriormente, se tiene que:

$$F = \frac{(21303,07 - 21223,89)/2}{21223,89/(23035 - 7)} = 42,95$$

El valor p del F=0.000000. De acuerdo con los resultados obtenidos, se rechaza la hipótesis nula, es decir, se debe seleccionar el modelo A,

pues al menos uno de los coeficientes que acompaña a las variables explicatorias adicionales es significativo.

Una forma alterna de efectuar la prueba anterior, y que nos conduce exactamente a los mismos resultados, es usar el comando **test**, explicado anteriormente, así:

Comando

```
test zona=sexo*zona=0
```

Modelos no anidados

Dos modelos son no anidados cuando ninguno de los dos está contenido en el otro, supongamos que se debe seleccionar entre el modelo A y el modelo B.

$$A: \ln(\text{salario mensual}) = \beta_0 + \beta_1 \cdot \text{Años de estudio}_i + \beta_2 \cdot \text{exper}_i + \beta_3 \cdot \text{exper}_i^2 + \alpha_1 \cdot \text{sexo}_i + \alpha_2 \cdot \text{zona}_i + \gamma \cdot (\text{sexo} * \text{zona})_i + u_i$$

$$B: \ln(\text{salario mensual}) = \beta_1 + \beta_2 \cdot \text{Años de estudio}_i + \beta_3 \cdot \text{exper}_i + \beta_4 \cdot \text{exper}_i^2 + \alpha_1 \cdot \text{sexo}_i + \alpha_2 \cdot (\text{sexo} * \text{Años de estudio})_i + u_i$$

Para seleccionar entre dos modelos no anidados que contienen la misma variable dependiente, pero diferente número de variables explicativas, se utilizan diversos criterios, como el R^2 ajustado, el criterio de información de Akaike y el criterio de información de Schwarz, los cuales serán tratados en la sección Criterios de selección de modelos.

Otra forma de seleccionar entre dos modelos es usando la prueba J de Davidson-MacKinnon. Para ilustrar su aplicación consideremos los dos modelos no anidados anteriores y procedamos siguiendo estos pasos:

1. Estime el modelo A por MCO y obtenga los valores de la variable dependiente estimados o ajustados a través del comando:

Comando

```
predict yhat,xb
```

2. Estime el modelo B agregando como variable explicatoria la variable dependiente ajustada del modelo A.
3. Revise, después de estimar B, si el coeficiente de la variable ajustada del modelo A incluida en B es significativo, de serlo, el modelo B no es correcto; en caso contrario, es correcto.
4. Para evaluar el modelo A repita el mismo procedimiento, es decir, estime el modelo B y obtenga los valores estimados o ajustados de su variable dependiente.
5. Estime el modelo A agregando como variable explicatoria la variable dependiente ajustada del modelo B.
6. Revise, después de estimar A, si el coeficiente de la variable ajustada del modelo B incluida en A es significativo, de serlo, el modelo A no es correcto; en caso contrario, es correcto.

El procedimiento anterior puede producir que ambos modelos sean correctos o que ambos modelos sean incorrectos, los resultados obtenidos utilizando la prueba J para la selección entre los dos modelos de interés concluyen que ninguno de los dos modelos es correcto.

4.3.3 Criterios de selección de modelos

Existen varios criterios para seleccionar modelos que compiten entre sí, la mayoría de dichos criterios tienen como filosofía minimizar la suma residual de cuadrados, imponiendo un castigo a aquellos modelos que teniendo la misma variable dependiente y el mismo número de observaciones incluyen un mayor número de variables explicatorias.

Los criterios a considerar, entre otros, son: R^2 ajustado, criterio de información Akaike (CIA) y criterio de información bayesiana de Schwarz (CIB).

R^2 ajustado

El R^2 ajustado se utiliza cuando se necesita comparar dos modelos que tienen la misma variable dependiente, pero diferente número de vari-

ables explicatorias. Este indicador de ajuste es mejor que el coeficiente de determinación R^2 , en términos de comparación de modelos, en la medida que tiene en cuenta el número de variables explicatorias, mientras que el coeficiente de determinación tradicional es una función no decreciente del número de variables explicatorias, es decir, cuantas más variables explicatorias, mayor es el R^2 . El R^2 ajustado se define como:

$$\overline{R^2} = 1 - (1 - R^2) \frac{n - 1}{n - k} \quad (4.4)$$

Criterio de información Akaike

La idea de imponer una penalización por anadir regresoras al modelo se desarrolló mas en el criterio de información de Akaike (CIA), el cual se define como:

$$CIA = \ln \left(\frac{SCR}{n} \right) + \frac{2k}{n} \quad (4.5)$$

Donde:

- k es el número de parámetros,
- SCR es la suma residual de cuadrado, y
- n es el número de observaciones.

El CIA impone una mayor penalización que R^2 por añadir regresoras. Al comparar dos o más modelos, se preferirá el que tenga el menor valor CIA.

Una ventaja del CIA es que resulta útil para comparar el desempeño de la predicción dentro y fuera de la muestra, y también ayuda a determinar la longitud del rezago óptimo en los modelos autorregresivos de orden p (Gujarati y Porter, 2010).

Criterio de información bayesiano

El criterio de información bayesiano (CIB), o de Schwarz, busca penalizar la incorporación de variables explicatorias en el modelo de regresión, y se define como:

$$CIB = \ln \left(\frac{SCR}{n} \right) + \frac{k}{n} * \ln(n) \quad (4.6)$$

Al igual que en CIA, mientras más pequeño sea el valor de CIB, mejor será el modelo. De igual manera que el CIA, el CIB también sirve para comparar el desempeño de la predicción dentro y fuera de la muestra, y determinar la longitud del rezago óptimo en los modelos autorregresivos de orden p (Gujarati y Porter, 2010).

El criterio de Schwarz es el más exigente por la inclusión de regresores en el modelo. En cualquier situación penaliza más que el resto. Schwarz aumenta su penalización y por ende en el infinito tiende a aumentar la probabilidad de elegir el modelo correcto. Por eso es el único de los tres criterios que es consistente.

En Stata, para seleccionar un modelo mediante el CIA o el CIB, primero se deben estimar los modelos y posteriormente digitar el comando **estat ic**, seleccionando el modelo que menor medida arroje en estos indicadores. De acuerdo con los criterios anteriores, entre los modelos no anidados A y B se debe seleccionar el modelo B, pues los resultados arrojados para el modelo A son: R^2 ajustado= 0.307, criterio de información de Akaike=2.756, criterio de información de Schwarz=2.759, mientras que para el modelo B son: R^2 ajustado= 0.316, criterio de información de Akaike=2.743 y criterio de información de Schwarz=2.745.

Análisis de regresión con series de tiempo

Temas Tratados

- 5.1 Estimación e interpretación de un modelo con datos de series de tiempo**
 - 5.1.1 Estadísticas de influencia**
- 5.2 Autocorrelación**
 - 5.2.1 ¿Cómo detectar autocorrelación?**
 - 5.2.2 ¿Cómo solucionar autocorrelación?**
- 5.3 Modelos econométricos dinámicos**
 - 5.3.1 Estimación de un modelo dinámico**
- 5.4 Test de causalidad**
- 5.5 Desestacionalización**
- 5.6 Raíz unitaria y cointegración univariada**
 - 5.6.1 Pruebas de raíz unitaria y cointegración univariada**

En este capítulo se inicia el análisis de regresión con datos de series de tiempo. Las razones para separar los modelos de regresión que utilizan datos de corte transversal de los que utilizan series de tiempo, se encuentran muy bien explicadas en (Wooldridge, 2010). Una de las razones es que los datos de corte transversal facilitan la comprensión de los análisis econométricos y sirven de referencia para la introducción al análisis econométrico con datos de series de tiempos, los cuales tienen en su estructura, por lo general, componentes de tendencia, estacionalidad, ciclicidad y aleatoriedad, volviendo más complejo el análisis de regresión.

5.1 Estimación e interpretación de un modelo con datos de series de tiempo

Para estudiar los modelos de regresión con datos en series de tiempo se parte del modelo paridad de poder adquisitivo (PPA) relativo, el cual establece que las variaciones en la tasa de cambio dependen de las variaciones en los índices de precios de Colombia y Estados Unidos. Adicionalmente, se introduce una variable *dummy* para capturar la entrada de grandes volúmenes de divisas al país, debido al aumento de la inversión extranjera a partir del año 2004. Dicha variable *dummy* toma el valor de 0 antes del 2004 y de 1 a partir del 2004. Los datos utilizados corresponden al periodo de 1961 al 2011. Los de inflación fueron recogidos por el Banco Mundial; mientras que los de variaciones de la tasa de cambio fueron recogidos por el Banco de La República.

El modelo planteado se presenta en la [ecuación \(5.1\)](#). Según la PPA relativa, un incremento en la tasa de inflación externa aumenta el poder adquisitivo de nuestra moneda y, por tanto, debe producirse una apreciación nominal de nuestra moneda frente al exterior. Un aumento de la tasa de inflación interna reduce el poder adquisitivo de nuestra moneda y, por tanto, debe producirse una depreciación nominal de nuestra moneda.

$$VTCN_t = \beta_1 + \beta_2 \pi_{EEUU}_t + \beta_3 \pi_{COL}_t + \beta_4 Dummy_t + U_t \quad (5.1)$$

Para estimar el modelo anterior en Stata se debe declarar que los datos son series de tiempo y la frecuencia de los mismos. En ocasiones, los datos son de frecuencia mensual o trimestral, y al importarlos desde Excel, la fecha se importa como una variable *string* o no numérica. Para solucionar este problema, una salida es generar una nueva variable para luego declararla como serie de tiempo. La forma de generar esta variable depende de la frecuencia de los datos, por lo tanto, si los datos son trimestrales, la generación debe realizarse de la siguiente manera:

Comando

```
generate t= tq(1978q1)+_n-1
```

La instrucción anterior está informando que se generó una variable *t* de frecuencia trimestral con información desde el primer trimestre de 1978. Si la información es mensual, lo único que se modifica es la *q* por *m*, así:

Comando

```
generate t= tm(1978m5)+_n-1
```

La instrucción anterior está informando que se generó una variable *t* de frecuencia mensual con información desde mayo de 1978.

Generada la variable *t* debe declararse como serie de tiempo siguiendo esta ruta:

Ventana

1. statistics → time → setup and utilities → declare data set to be series time data → ↪
2. En time variable señale la variable generada así como su frecuencia → ok.

Una forma alterna de declarar la variable generada es a través del comando *tsset*, nombre de la variable generada o de la serie de tiempo coma (,) frecuencia de los datos. Para el caso trimestral por comando sería:

Comando

```
tsset t, quarterly
```

Después de declarada la serie se estima el modelo común y corriente.

La [Figura 5.1](#) presenta la estimación obtenida para el modelo planteado anteriormente

. regress VTCN nEEUU nCOL Dummy						
Source	SS	df	MS	Number of obs	=	51
Model	3581.85152	3	1193.95051	F(3, 47)	=	16.10
Residual	3485.69137	47	74.1636461	Prob > F	=	0.0000
Total	7067.54289	50	141.350858	R-squared	=	0.5068
				Adj R-squared	=	0.4753
				Root MSE	=	8.6118
VTCN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nEEUU	-1.146997	.4887985	-2.35	0.023	-2.130333	-.1636612
nCOL	.6938138	.1838689	3.77	0.000	.3239174	1.06371
Dummy	-11.83137	4.015077	-2.95	0.005	-19.90866	-3.754076
_cons	7.84807	3.382185	2.32	0.025	1.043992	14.65215

Figura 5.1: Estimación MCO con series de tiempo.

Los resultados arrojados por el Stata muestran una estructura similar a la presentada cuando los datos eran de corte transversal. Por otro lado, los valores P asociados a las variables explicatorias son inferiores al 5%, lo que implica que la variación de precios en Colombia y Estados Unidos, así como la variable *dummy*, explican la variación de la tasa de cambio nominal.

En términos de las interpretaciones de los coeficientes estimados y del *R*², se tiene:

- $\beta_1 = -1,14$ = ante un aumento de un punto porcentual en la inflación de los Estados Unidos, la variación en la tasa de cambio esperada disminuye 1.14 puntos porcentuales, *ceteris paribus*.
- $\beta_2 = 0,69$ = ante un aumento de un punto porcentual en la inflación colombiana, la variación en la tasa de cambio esperada aumenta 0.69 puntos porcentuales, *ceteris paribus*.
- $\beta_3 = -11,83$ = después del 2004, la tasa de cambio esperada disminuye 11.83 puntos porcentuales frente al periodo antes del 2004.
- $R^2 = 0,5068$ = el 50.68% de la variación total de las variaciones porcentuales de la tasa de cambio es explicada por el modelo.

5.1.1 Estadísticas de influencia

Existen diversas pruebas que permiten verificar si hay valores atípicos y saber cuál es la influencia de estos sobre los valores de las estimaciones obtenidas; entre estas se destacan los residuos studentizados, los DFFITS y los DFBETAS. Estas pruebas se pueden hacer también en modelos con datos de corte transversal.

Para realizar estas pruebas se debe estimar primero el modelo planteado y posteriormente seguir esta ruta:

Ventana

View → Stability Diagnostics → Influence Statistics → escoger la prueba a realizar.

Residuos estudiantizados

Los residuos estudiantizados permiten detectar observaciones atípicas, aunque presentan el mismo problema de los residuos MCO, es decir, son heterocedásticos y autocorrelacionados, aún si los errores verdaderos tienen varianza común y son serialmente independientes. Cuando los residuos estudiantizados son superiores a 2 en valor absoluto se considera que la observación asociada es atípica, lo que implica un cuidado especial en el análisis. Por otro lado, la presencia de muchos valores atípicos puede ocasionar el no cumplimiento de normalidad en los errores. La [Figura 5.2](#) muestra los residuos estudiantizados.

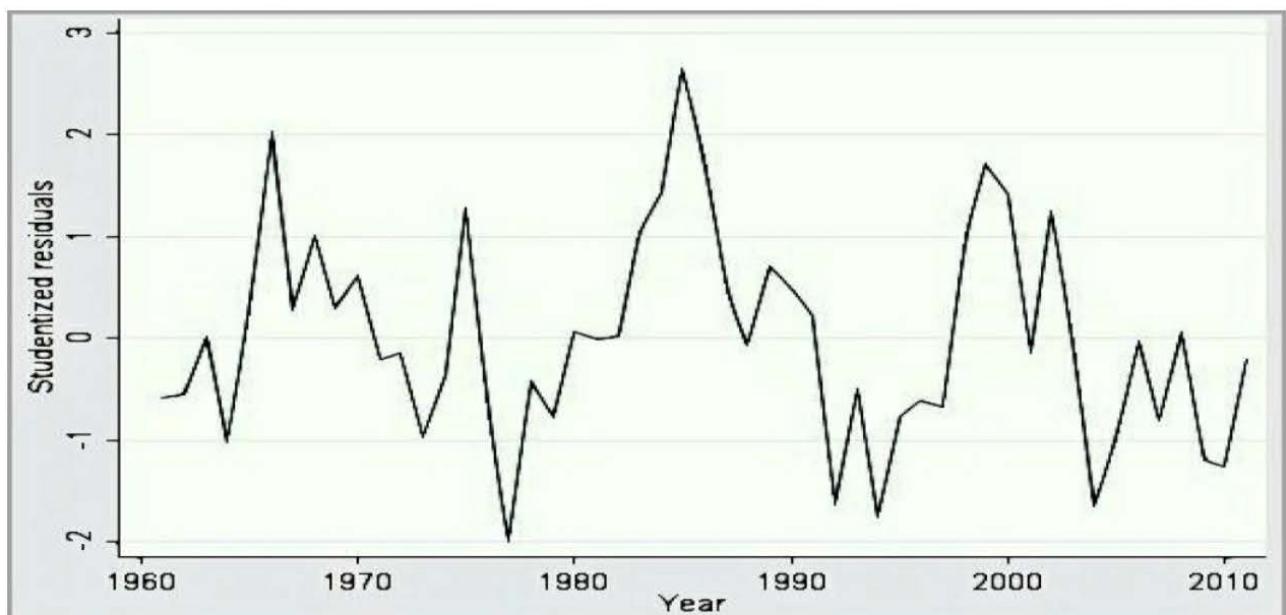


Figura 5.2: Residuos estudiantizados.

Para obtener las gráficas de los residuos estudiantizados como se presenta en la [Figura 5.2](#), se estima el modelo en consideración y se sigue esta ruta:

Ventana

statistics → postestimation → predictions → residuals → → active studentized residual → asignar un nombre a sus residuos → ok.

Obtenida la serie de los residuos estudiantizados se procede a graficarla así:

Ventana

graphics → twoway graph → → create/line/en → señalar el nombre de los residuos estudiantizados → en x variable señale su variable tiempo → Ok.

Por comandos sería:

Comando

Para las gráficas por línea de comando sería:

Comando

```
twoway (line nombre_de_los_residuos_studentizados t) donde t es el nombre de la variable tiempo definido cuando se declaró la serie de tiempo
```

DFFITS

DFFITS es una medida que se construye a partir de los residuos estudentizados y permite detectar valores atípicos y observaciones influentes, el resultado es mostrado en la [Figura 5.3](#). Una observación es influente si se cumple la condición de la [ecuación \(5.2\)](#).

$$DFFITS(i) > 2\sqrt{\frac{k}{n}} \quad (5.2)$$

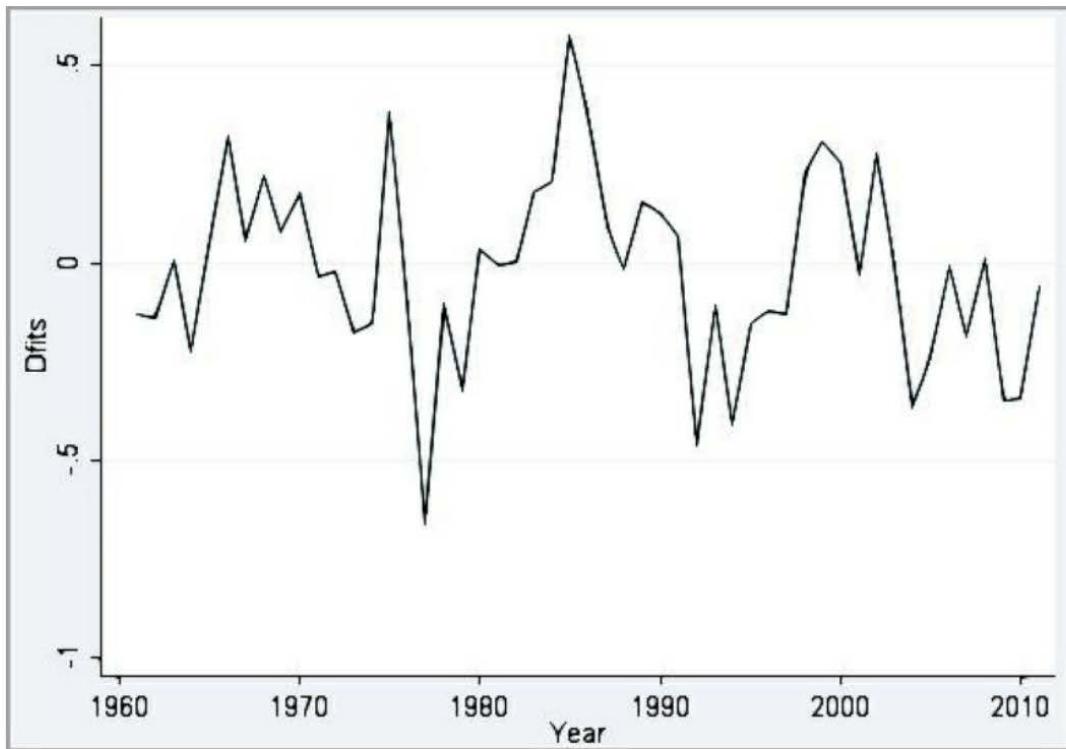


Figura 5.3: DFFITS.

Para obtener las gráficas de los DFFITS, como se presenta en la [Figura 5.3](#), se estima el modelo en consideración y se sigue esta ruta:

Ventana

statistics → postestimation → predictions → residuals → active Dfits → asignarle un nombre a su Dfits → Ok.

Obtenidas las series de los DFFITS se procede a graficarlas así:

Ventana

graphic → twoway graph → create → line → en Y variable señalar el nombre de los DFITS → en X variable señalar su variable tiempo → Ok.

Usando la línea de comando sería:

Comando

```
predict (nombre de los dfits), dfits
```

Para generar las gráficas usando comando sería:

Comando

```
twoway (line nombre_de_los_dfits t), donde t es el nombre de la variable tiempo definida cuando se declaró la serie de tiempo
```

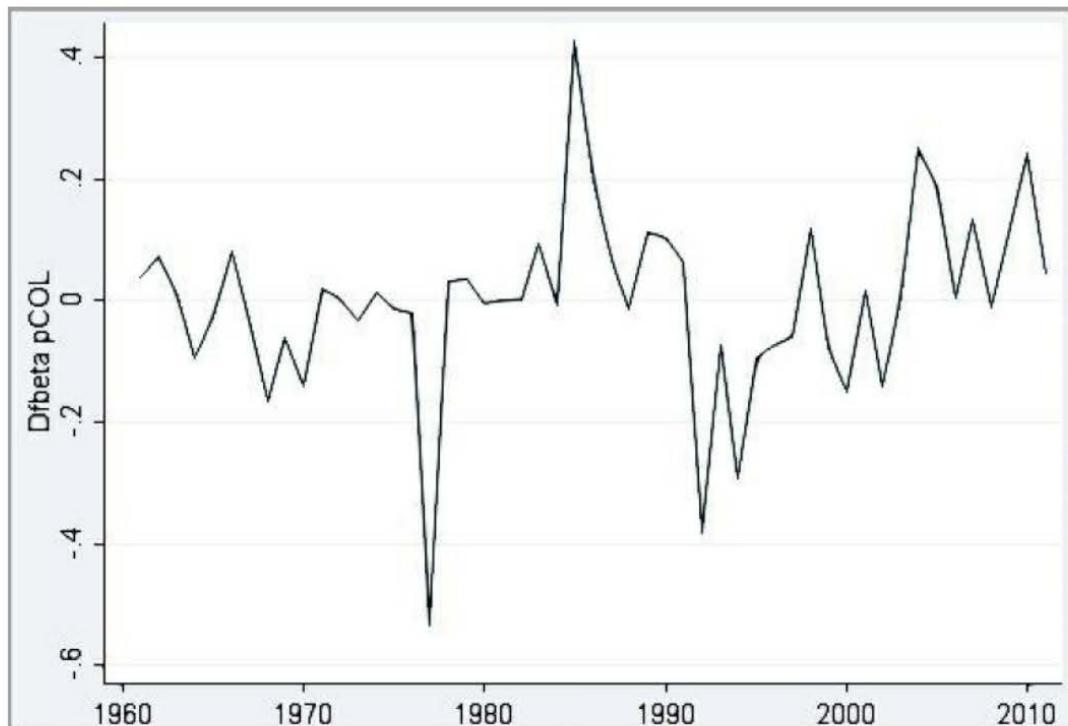
DFBETAS

DFBETAS es una medida utilizada para determinar si existen observaciones que impactan o influyen de manera especial sobre los valores de los parámetros estimados. Cuando el valor de DFBETAS es mayor en valor absoluto a $2/\sqrt{n}$, indica que existe tal influencia. Los DFBETAS proporcionan esta medida para cada coeficiente de regresión estimado. En Stata se calcularán los DFBETAS de los coeficientes de regresión que acompañan a las variables explicatorias. Ver [Figuras 5.4a](#) y [5.4b](#).

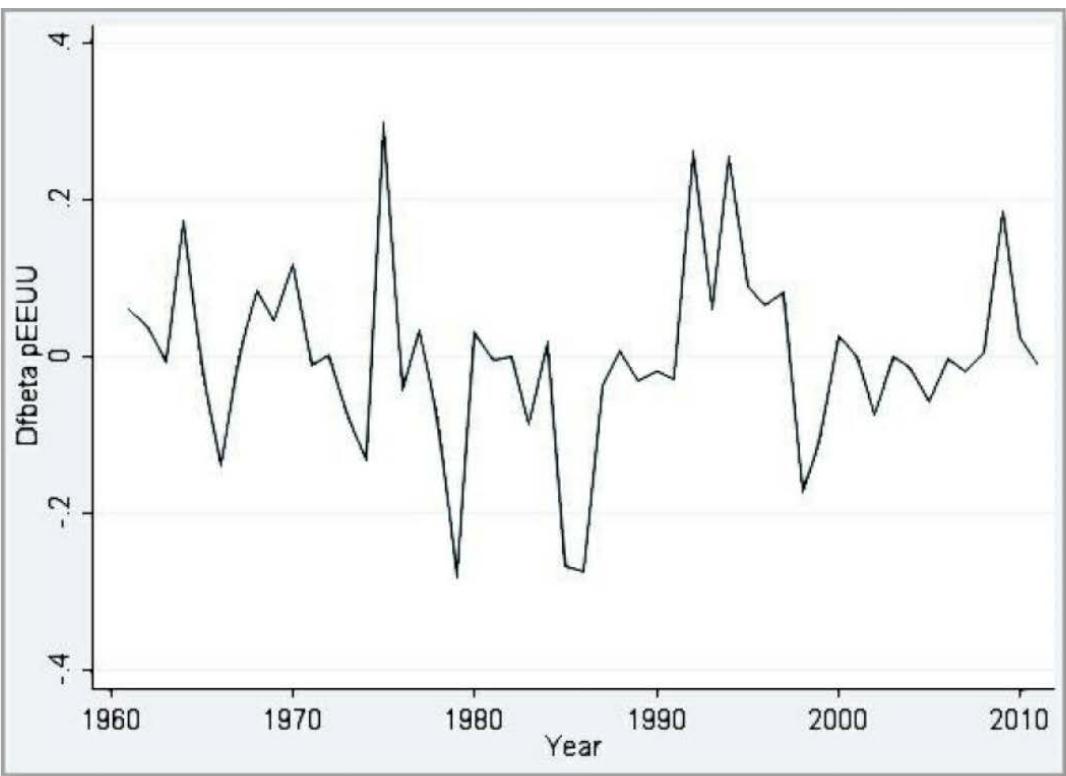
Para obtener las gráficas de los DFBETAS, como se presentan en las [Figuras 5.4a](#) y [5.4b](#), se estima el modelo en consideración y se sigue esta ruta:

Ventana

statistics → linear model and related → regression diagnostics → Dfbeta → (señalar las variables explicatorias a las cuales se les quiere obtener el Dfbeta) → Ok.



(a) DFBETA Colombia.



(b) DFBETA Estados Unidos.

Obtenidas las series de DFBETAS para cada coeficiente de regresión, se procede a graficarlas así:

Ventana

```
graphics → twoway graph → create → line → en Y variable señalar el DFBETA a graficar → en X variable señalar su variable tiempo → Ok.
```

El proceso realizado por ventanas se puede llevar a cabo por comandos de la siguiente forma:

Comando

```
dfbeta (lista de variables explicatorias)
```

Para generar las gráficas por línea de comando se hace lo siguiente:

Comando

```
twoway (line dfbeta 1 t)
```

Donde **DFBETA_1** es el nombre asignado por defecto (*default*) al coeficiente de la primera variable explicatoria, y *t* es la variable tiempo definida cuando se declaró la serie de tiempo. Por lo tanto, Stata produce una gráfica de DFBETA por cada variable explicatoria.

5.2 Autocorrelación

El modelo supone que las perturbaciones o errores del modelo no están correlacionados, es decir, $\text{Cov}(u_t, u_s) = 0$, sin embargo, en estudios que utilizan series de tiempo, por lo general, dicho supuesto se rompe, es decir, $\text{Cov}(U_t, U_s) \neq 0$, esto implica que los estimadores **MCO** dejen de ser **MELI**, dado que ya no poseen varianza mínima.

Existen diversos factores que pueden generar autocorrelación, dentro de estos se pueden mencionar: la inercia propia de las variables económicas, la omisión de variables, la manipulación de los datos y la falta de estacionariedad en las series.

5.2.1 ¿Cómo detectar autocorrelación?

Podemos usar diferentes métodos para detectar la autocorrelación, entre ellos, el método gráfico, el test de Durbin Watson y el test de Breusch-Godfrey

El método gráfico

Es un método informal que consiste en elaborar un gráfico de \hat{u}_{t-1} versus \hat{u}_t . Si existe un patrón sistemático de comportamiento en dicha gráfica, se presume que hay indicios de autocorrelación en los errores. Para generar los residuos rezagados en un periodo, Stata hace uso del operador L, por lo tanto, $\hat{u}_{t-1} = L. \hat{u}_t$.

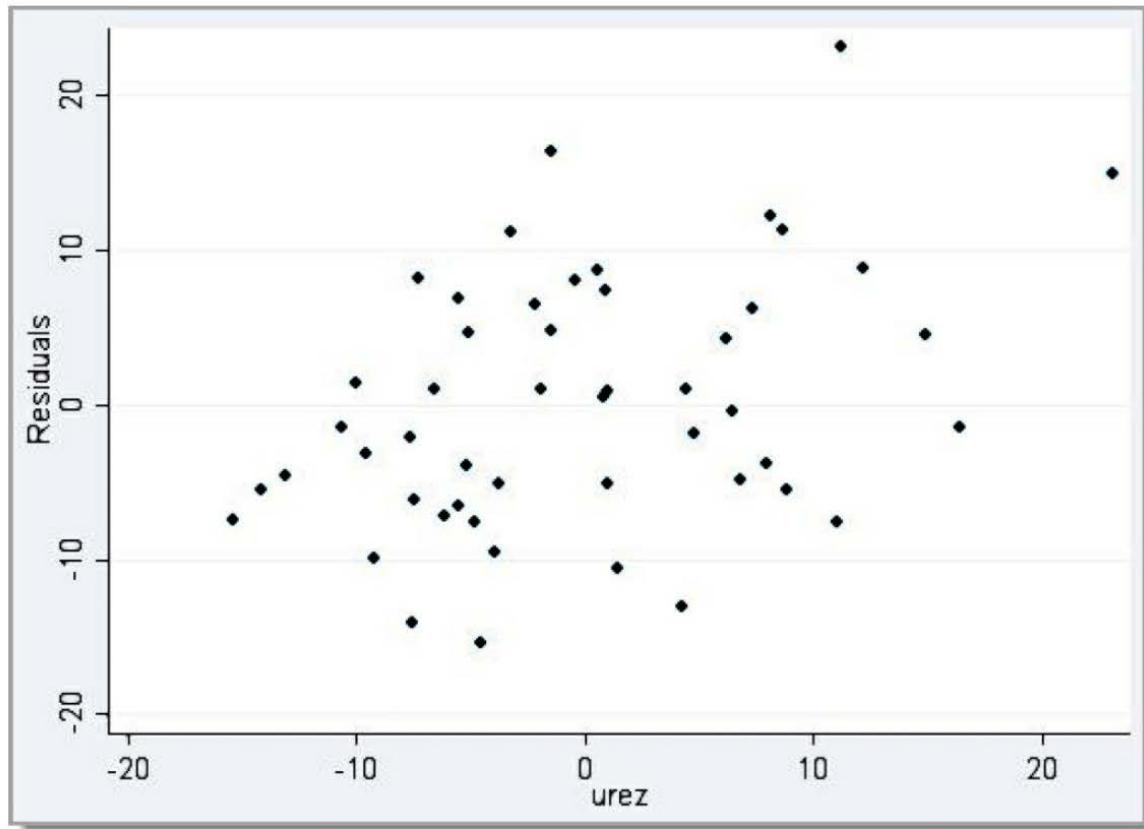


Figura 5.5: Gráfica de \hat{u}_{t-1} versus \hat{u}_t .

La [Figura 5.5](#) de residuos versus los residuos rezagados muestra un comportamiento sistemático, por consiguiente, se puede pensar que existe autocorrelación en los errores del modelo. Obsérvese que a medida que los residuos rezagados aumentan, los residuos también aumentan. No obstante, el método gráfico es muy subjetivo y por ello debe ser validado con tests formales, como los que se presentan a continuación.

Prueba de Durbin Watson

Esta prueba es aplicable bajo las siguientes condiciones: el modelo estimado tiene intercepto; los errores siguen una distribución normal; las variables explicativas son fijas o, si son aleatorias, no están correlacionadas con el error; la variable dependiente no puede figurar como explicatoria rezagada; no hay observaciones faltantes y se asume que los errores siguen un proceso autorregresivo de orden 1, es decir:

$$u_t = \rho u_{t-1} + e_t$$

siendo e_t un proceso ruido blanco, lo que quiere decir que su media es 0 para cualquier t, su varianza es constante para todo t y no presenta autocorrelación.

$$E(e_t) = 0, \quad V(e_t) = \sigma^2, \quad Cov(e_t, e_s) = 0, \quad t \neq s$$

La hipótesis nula en esta prueba es H_0 : no existe autocorrelación en los errores y H_1 : existe autocorrelación en los errores.

El estadístico de prueba Durbin Watson se define en la [ecuación \(5.3\)](#).

$$d = \frac{\sum_2^n (\widehat{u}_t - \widehat{u}_{t-1})^2}{\sum_1^n \widehat{u}_t^2} \approx 2(1 - \widehat{\rho}) \quad (5.3)$$

El estadístico Durbin Watson toma valores entre 0 y 4, a medida que dicho estadístico se aproxime a 0 indica presencia de autocorrelación positiva, si se aproxima a 4 indica presencia de autocorrelación negativa, y en la medida que se aproxima a 2 indica ausencia de autocorrelación. En Stata, el comando para obtener el estadístico de Durbin Watson es **estat dwatson**.

regress VTCN nEEUU nCOL Dummy						
Source	SS	df	MS	Number of obs	=	51
Model	3581.85152	3	1193.95051	F(3, 47)	=	16.10
Residual	3485.69137	47	74.1636461	Prob > F	=	0.0000
Total	7067.54289	50	141.350858	R-squared	=	0.5068
				Adj R-squared	=	0.4753
				Root MSE	=	8.6118
VTCN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nEEUU	-1.146997	.4887985	-2.35	0.023	-2.130333	-.1636612
nCOL	.6938138	.1838689	3.77	0.000	.3239174	1.06371
Dummy	-11.83137	4.015077	-2.95	0.005	-19.90866	-3.754076
_cons	7.84807	3.382185	2.32	0.025	1.043992	14.65215

estat dwatson
Durbin-Watson d-statistic(4. 51) = 1.192125

Figura 5.6: Test de Durbin-Watson.

La Figura 5.6 muestra el resultado del estadístico de prueba para el caso de estudio, el cual arroja un valor para el estadístico Durbin Watson de 1.192, lo que sugiere presencia de autocorrelación positiva, pues al contrastar dicho valor con los valores tabulados para tres variables explicativas, un tamaño de muestra de 51 Observaciones y un nivel de significancia del 5%, **dl=1.427** y **du=1.675**, se concluye que existe autocorrelación en los errores del modelo.

Prueba de Breusch-Godfrey

Esta prueba es más general que la de Durbin Watson, ya que asume que el proceso generador de datos puede ser un proceso autorregresivo de orden **P, AR(p)**,

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + \rho_3 U_{t-3} + \cdots + \rho_p U_{t-p} + e_t \quad (5.4)$$

O un proceso de media móvil de orden **q**, es decir:

$$U_t = \theta_1 e_{t-1} + \theta_2 e_{t-2} + \theta_3 e_{t-3} + \cdots + \theta_q e_{t-q} + e_t \quad (5.5)$$

donde e_t es un proceso ruido blanco. Una de las desventajas de la prueba de Breusch-Gofrey es que no determina el orden del rezago **P** en el proceso autorregresivo **AR(p)**, ni el **q** en el proceso de media móvil **MA(q)**, teniendo que recurrir para ello a los criterios de información como Akaike y Schwartz, ya presentados.

La hipótesis nula en esta prueba es:

$$H_0 : \text{no existe autocorrelación en los errores.}$$

$$H_1 : \text{existe autocorrelación en los errores.}$$

La ruta para realizar la prueba de Breusch-Gofrey, después de haber estimado el modelo, es:

Ventana

statistics → postestimation → reports and statistics → buscar la casilla (Breusch-Gofrey) → activar la celda para fijar el número de rezagos a ser probados → Ok.

Usando línea de comando sería:

Comando

```
estat bgodfrey, lags(rezagos a ser probados)
```

De acuerdo con los resultados presentados en la [Figura 5.7](#), se observa que para rezagos de orden 2, 3 y 4, los valores **p** asociados al estadístico de prueba chi cuadrado son inferiores al 5%; por lo tanto, se concluye que existe autocorrelación en los errores del modelo.

. regress VTCN nEEUU nCOL Dummy						
Source	SS	df	MS	Number of obs	=	51
Model	3581.85152	3	1193.95051	F(3, 47)	=	16.10
Residual	3485.69137	47	74.1636461	Prob > F	=	0.0000
Total	7067.54289	50	141.350858	R-squared	=	0.5068
				Adj R-squared	=	0.4753
				Root MSE	=	8.6118
. estat bgodfrey, lags(2 3 4)						
Breusch-Godfrey LM test for autocorrelation						
lags(p)	chi2	df		Prob > chi2		
2	8.175	2		0.0168		
3	8.197	3		0.0421		
4	9.075	4		0.0592		
H0: no serial correlation						

Figura 5.7: Test de Breusch-Gofrey.

5.2.2 ¿Cómo solucionar autocorrelación?

Identificado el problema de autocorrelación, el investigador debe solucionarlo, y para ello debe suponer el proceso generador de errores a partir del cual debe transformar adecuadamente el modelo para eliminar dicha autocorrelación. Asumiendo que el proceso generador de errores es un

AR(1), la transformación adecuada es realizar cuasi primeras diferencias donde el estimador de la correlación entre los errores se puede estimar por el método de Durbin Watson o el método iterativo de Cochrane Orcutt. En este manual usaremos el método iterativo de Cochrane Orcutt, dado que es el de mayor uso.

Después de haber estimado el modelo que presenta el problema, la ruta a seguir es:

Ventana

statistics → Times series → prals-Winsten → defina su variable dependiente y sus variables independientes o explicatorias → activar la casilla Cochrane- Orcutt transformation → Ok.

Podemos obtener el mismo proceso usando la siguiente línea de comando:

Comando

```
prals VTCN nEEUU nCOL Dummy, rtype(regress) corc
```

De acuerdo con los resultados que se presentan en la [Figura 5.8](#), se observa que el estadístico de Durbin-Watson inicial de 1.19 pasa a 2.05 en el modelo transformado con cuasi diferencias, solucionando de esta forma el problema de autocorrelación presentado, garantizando, por consiguiente, que los estimadores obtenidos son MELI.

```
Iteration 3: rho = 0.4068
Iteration 4: rho = 0.4070
Iteration 5: rho = 0.4070
Iteration 6: rho = 0.4070
Iteration 7: rho = 0.4070
```

Cochrane-Orcutt AR(1) regression -- iterated estimates

Source	SS	df	MS	Number of obs	=	50
Model	1820.05176	3	606.683918	F(3, 46)	=	9.76
Residual	2858.36061	46	62.1382741	Prob > F	=	0.0000
Total	4678.41236	49	95.4778034	R-squared	=	0.3890
				Adj R-squared	=	0.3492
				Root MSE	=	7.8828

VICN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
nEEUU	-.9511525	.6035923	-1.58	0.122	-2.166121 .2638158
nCOL	.7182905	.2058138	3.49	0.001	.3040089 1.132572
Dummy	-11.61059	5.187634	-2.24	0.030	-22.05276 -1.168426
_cons	6.967248	4.348073	1.60	0.116	-1.784969 15.71946
rho	.4069789				

Durbin-Watson statistic (original) 1.192125

Durbin-Watson statistic (transformed) 2.053877

Figura 5.8: Solución de autocorrelación.

Hasta el momento se han considerado modelos estáticos los cuales presentan la siguiente estructura:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t \quad (5.6)$$

En dichos modelos se supone que la variable dependiente y las explicatorias se relacionan contemporáneamente, es decir, que el impacto que tiene cada variable explicatoria sobre la variable dependiente, medido a través de los coeficientes de regresión, es instantáneo. En otras palabras, un cambio en la variable explicatoria X_2 en el periodo t impacta a la variable dependiente Y en el mismo periodo de tiempo t , lo cual, desde el punto de vista de la política económica, no es cierto, pues el efecto total de una política, por ejemplo, un incremento en la tasa de interés para controlar inflación no se deja sentir en su totalidad en el mismo periodo de tiempo, por lo general, dicho efecto se manifiesta parcialmente a lo largo de cierto número de periodos. Esta limitación de los modelos estáticos, desde el punto de vista aplicado, plantea la necesidad de incorporar el tiempo en el diseño del modelo, que refleje la dinámica de las variables económicas, apareciendo así los modelos dinámicos, los cuales se pueden clasificar en modelos de rezagos distribuidos infinitos, modelos de rezagos distribuidos finitos y modelos autorregresivos, como se describen a continuación.

1. **Modelos de rezagos distribuidos infinitos.** Un modelo de rezagos distribuidos infinitos tiene la estructura que se presenta en la [ecuación \(5.7\)](#).

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \dots + U_t \quad (5.7)$$

En este modelo, X tiene dos tipos de impacto sobre Y , uno de corto plazo medido por β_0 , y otro de largo plazo medido por la suma de los impactos parciales $\sum_0^\infty \beta_i$.

2. **Modelos de rezagos distribuidos finitos.** Un modelo de rezagos distribuidos finitos tiene la estructura que se presenta en la [ecuación \(5.8\)](#).

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + U_t \quad (5.8)$$

En este modelo se considera que los impactos de X sobre Y se distribuyen a lo largo del tiempo en un impacto instantáneo medido por β_0 , más tres impactos parciales los cuales se dejan sentir uno, dos y tres periodos después. β_2 mide el impacto que tiene X sobre Y dos periodos después.

3. **Modelos autorregresivos.** Un modelo autorregresivo tiene la estructura que se presenta en la [ecuación \(5.9\)](#).

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 Y_{t-1} + U_t \quad (5.9)$$

En este modelo, se considera la verdadera dinámica, donde uno de los factores que explica el comportamiento de Y en el momento actual es su propio pasado.

5.3.1 Estimación de un modelo dinámico

Para ilustrar la estimación de un modelo dinámico, se considera un modelo autorregresivo donde las variaciones en la tasa de cambio depende tanto de las variaciones en los precios tanto de Colombia como de los Estados Unidos y de las variaciones en la tasa de cambio del periodo anterior. Indudablemente, asumir este modelo cuando la frecuencia de los datos es anual no tiene mucho sentido, no obstante, se asumirá de esta forma para poder ilustrar la estimación de dicho modelo. Para la estimación de estos modelos se utilizará la misma base de datos con sus fuentes empleada en la estimación de un modelo con datos de series de tiempo ([sección 5.1](#)). La creación de variables rezagadas en Stata es fácil, simplemente se genera una nueva variable que contendrá la variable rezagada anteponiendo la letra **I**. a la variable que se pretende rezagar. Si el número de rezagos es 2, se antepone **I2**, si el número de rezagos es 3 se antepone **I3**, y así sucesivamente, por ejemplo: si se quiere rezagar un periodo la variable **VTCN** se procede de la siguiente manera:

Comando

```
generate VTCNREZ=I.VTCN
```

Así la variable **VTCNREZ** contendrá a la variable **VTCN** rezagada un periodo. Si se requiere rezagar dos periodos, simplemente se coloca **I2.VTCN**, y así sucesivamente.

Modelo autorregresivo

La [ecuación \(5.10\)](#) describe la estructura del modelo autorregresivo considerado para la variación de la tasa de cambio.

$$VTCN_t = \beta_0 + \beta_1 \pi_{COL_t} + \beta_2 \pi_{EEUU_t} + \lambda_1 VTCN_{t-1} + U_t \quad (5.10)$$

La [Figura 5.9](#) muestra la estimación mínima cuadrática del modelo anterior.

reg VTCN pCOL pEEUU rezVTCN

Source	SS	df	MS	Number of obs	=	50
Model	3644.83481	3	1214.94494	F(3, 46)	=	16.77
Residual	3332.6794	46	72.4495521	Prob > F	=	0.0000
Total	6977.51421	49	142.398249	R-squared	=	0.5224
				Adj R-squared	=	0.4912
				Root MSE	=	8.5117

VTCN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pCOL	.6703335	.180573	3.71	0.001	.3068589 1.033808
pEEUU	-.6829236	.5150782	-1.33	0.191	-1.719722 .353875
rezVTCN	.3903272	.1199454	3.25	0.002	.1488897 .6317647
_cons	-.3930213	2.768203	-0.14	0.888	-5.965126 5.179083

Figura 5.9: Estimación modelo dinámico.

Un cuidado que se debe tener con este tipo de modelos en su estimación, es la presencia de autocorrelación pura en los errores, pues en este caso los estimadores MCO son sesgados e inconsistentes. Cuando la autocorrelación no es pura sino resultante de una mala especificación como omisión de variables relevantes, se debe resolver primero el problema de especificación, lo cual posiblemente generará que el término de error no presente autocorrelación y, por tanto, los estimadores obtenidos por MCO en grandes muestras serán insesgados y consistentes.

Existen dos maneras de determinar autocorrelación en los modelos autorregresivos: el **h de Durbin** y el test de **Breusch Gofrey**. El h de Durbin viene dado por la [ecuación \(5.11\)](#):

$$h = \hat{\rho}^2 \sqrt{\frac{n}{1 - nV(\hat{\alpha})}} \quad (5.11)$$

Donde:

- n es el tamaño de la muestra,
- $V(\hat{\alpha})$ es la varianza del coeficiente de la variable dependiente que aparece como explicatoria rezagada un periodo, y
- $\hat{\rho}$ es la estimación del coeficiente de correlación entre el error en el momento t y el error rezagado en un periodo.

La estimación del coeficiente de correlación se encuentra a partir del Durbin Watson arrojado por el paquete, pues sabemos que $d \approx 2(1 - \hat{\rho})$, entonces $\hat{\rho} = 1 - d/2$. Bajo la hipótesis nula $p=0$, h se distribuye asintóticamente normal, entonces si el h calculado cae entre -1.96 y 1.96 no se rechaza la hipótesis de no autocorrelación entre los errores a un nivel de significancia del 5%, en caso contrario, se rechaza. Para nuestro caso el h calculado es:

$$\hat{\rho} = 1 - 1,9389/2 = 0,03055$$

$$h = 0,03055^2 \sqrt{\frac{50}{1 - 50 * 0,0143}} = 0,404$$

De acuerdo con este resultado, los errores del modelo dinámico considerado no presentan autocorrelación, pues dicho valor se encuentra entre -1.96 y 1.96 .

La [Figura 5.10](#) presenta los resultados arrojados por Stata para el h de Durbin a través del comando `durbina`.

Model	3644.83481	3	1214.94494	F(3, 46)	=	16.77
Residual	3332.67939	46	72.4495521	Prob > F	=	0.0000
Total	6977.51421	49	142.398249	R-squared	=	0.5224
				Adj R-squared	=	0.4912
				Root MSE	=	8.5117

VICN	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
nEEUU	-.6829236	.5150782	-1.33	0.191	-1.719722 .353875
nCOL	.6703335	.180573	3.71	0.001	.3068589 1.033808
VICN					
L1.	.3903272	.1199454	3.25	0.002	.1488897 .6317647
_cons	-.3930213	2.768203	-0.14	0.888	-5.965126 5.179083

. durbina

Durbin's alternative test for autocorrelation

lags (p)	chi2	df	Prob > chi2
1	0.099	1	0.7528

H0: no serial correlation

Figura 5.10: Test h de Durbin.

Como lo plantea (Gujarati y Porter, 2010), la prueba h de Durbin presenta básicamente dos limitaciones: la primera es que es posible obtener un radicando negativo, en cuyo caso no es posible aplicarla; y la segunda, que es un test para muestras grandes, por tanto, en muestras pequeñas es cuestionable, de modo que se debe utilizar Breusch-Gofrey.

5.4 Test de causalidad

En ocasiones estamos interesados en determinar la dirección de causalidad entre dos variables, entendida como la información que le proporciona una variable a otra para mejorar su predicción y viceversa. En econometría de las series de tiempo se ha desarrollado la prueba de causalidad de Granger, la cual permite determinar la dirección de la causalidad, es decir, si **X** causa a **Y**, si **Y** causa a **X**, si ambas se causan mutuamente o si son independientes.

La prueba consiste en expresar a la variable dependiente **Y** en función de sus valores pasados y de los valores pasados de la variable independiente **X**, y de expresar la variable independiente **X** en función de sus valores pasados y los valores pasados de la variable dependiente **Y**, como se presentan en las [ecuaciones \(5.12\)](#) y [\(5.13\)](#):

$$Y_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_n X_{t-n} \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_n Y_{t-n} + U_{1t} \quad (5.12)$$

$$X_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_n Y_{t-n} \lambda_1 X_{t-1} + \lambda_2 X_{t-2} + \cdots + \lambda_n X_{t-n} + U_{2t} \quad (5.13)$$

Las dos hipótesis nulas que se consideran, según Granger, son las siguientes:

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$$

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_n = 0$$

- Si al usar el estadístico F de mínimos cuadrados restringidos se rechaza la primera hipótesis nula, pero no se rechaza la segunda hipótesis nula, entonces X causa a Y.
- Si al usar el estadístico F de mínimos cuadrados restringidos no se rechaza la primera hipótesis nula, pero se rechaza la segunda hipótesis nula, entonces Y causa a X.
- Si al usar el estadístico F de mínimos cuadrados restringidos se rechaza la primera hipótesis nula, y se rechaza la segunda hipótesis nula, entonces las dos variables se causan mutuamente.
- Si al usar el estadístico F de mínimos cuadrados restringidos no se rechaza la primera hipótesis nula y no se rechaza la segunda hipótesis nula, entonces X y Y son independientes.

Para ilustrar el test de Granger, supongamos que se desea determinar la dirección de la causalidad según Granger entre la inflación en Colombia y la variación de la tasa de cambio, y para ello se tienen en cuenta solo dos rezagos en la prueba. Las hipótesis a considerar son:

- H_0 = la inflación de Colombia no causa la variación de la tasa de cambio.
- H_a = la inflación de Colombia causa la variación de la tasa de cambio.

Dada la interconexión entre las variables económicas, el test de causalidad de Granger que ejecuta Stata considera las ecuaciones como un sistema de ecuaciones múltiples, por esta razón, los resultados arrojados por otro paquete estadístico, como por ejemplo el Eviews, no coinciden numéricamente, aunque sí en las decisiones finales alrededor de la causalidad. Para realizar este test por medio de Stata se sigue esta ruta:

Ventana

statistics → multivariate time series → vector autoregression → Definir las variables a las cuales se les quiere adelantar el test de causalidad (el paquete por default trae dos rezagos para cada variable) → Ok.

La ruta anterior por ventanas nos lleva a la estimación de un VAR(2), y se continúa con la ruta:

Ventana

statistics → multivariate time series → VAR diagnostics and tests → Granger causality tests → Ok.

. vargranger

Granger causality Wald tests

Equation	Excluded	F	df	df_r	Prob > F
VTCN	pCOL	1.9426	2	44	0.1554
VTCN	ALL	1.9426	2	44	0.1554
pCOL	VTCN	.16693	2	44	0.8468
pCOL	ALL	.16693	2	44	0.8468

Figura 5.11: Test de causalidad de Granger

Los resultados arrojados por el Stata y presentados en la [Figura 5.11](#) muestran que los valores P en ambas hipótesis son mayores que un alfa del 5%, por tanto, no se rechazan las hipótesis nulas, lo que implica que ambas variables consideradas son independientes, según Granger; es decir, que la inflación en Colombia no causa la variación de la tasa de cambio y que la variación de la tasa de cambio no causa la inflación de Colombia.

Es importante, como lo sugiere (Gujarati y Porter, 2010), tener las siguientes consideraciones al aplicar el test de causalidad de Granger:

- El test no determina la longitud del rezago que se debe considerar en la prueba, por tanto, eso se debe decidir con base en los criterios de

información ya tratados.

- Las series involucradas en la prueba deben ser estacionarias en el sentido débil, en covarianza o de segundo orden, entendida como aquellas series cuya media y varianza son constantes en el tiempo, y la covarianza depende de la longitud del rezago que las separa. En la siguiente sección se tratará el tema de estacionariedad débil o de segundo orden.
- Los errores en cada ecuación considerada no deben presentar autocorrelación.
- Dada la interconexión entre las variables económicas, el test de causalidad de Granger se debería ejecutar viendo las ecuaciones como un sistema de ecuaciones múltiples, como se desarrolló en esta sección.

5.5 Desestacionalización

Esta sección sigue fundamentalmente el texto de (Gujarati y Porter, 2010), y trata el problema de desestacionalización de una serie de tiempo cuya frecuencia de datos es mensual o trimestral, pues son estas series las que frecuentemente presentan efecto estacional, como por ejemplo, las ventas de vestuario en el comercio, las cuales presentan comportamientos sistemáticos todos los años, así: en el primer trimestre bajan las ventas, en el segundo suben un poco, en el tercero vuelven a bajar un poco y en el cuarto suben por incluir la época de Navidad. Para eliminar el efecto estacional existen diversos métodos, como **Census-X12**, **Census X-13**, **tramos**, entre otros. En este manual se considera el uso de las variables **dummy**, aunque se debe tener en cuenta que este método es válido si se asume que las componentes de la serie (tendencia, ciclo, estacional e irregular) son aditivas. La eliminación del efecto estacional se denomina desestacionalización o ajuste estacional.

Para ilustrar la técnica de desestacionalización usando variables dummy, se utiliza la tabla 9.4 sobre la venta de refrigeradores (en miles), medida con frecuencia trimestral para el período 1978.1-1985.4, del archivo de datos de (Gujarati y Porter, 2010) ([Figura 5.12](#)).

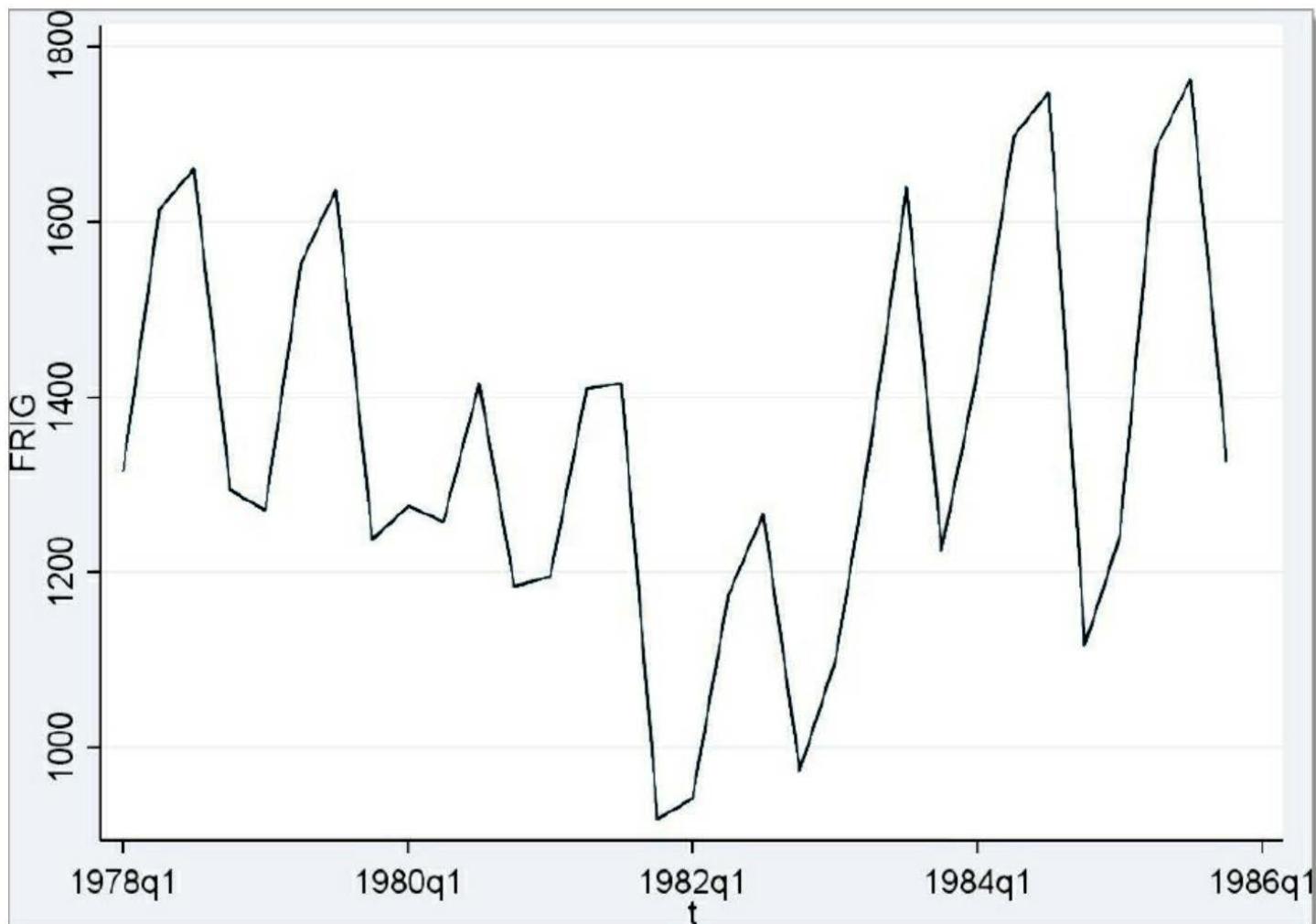


Figura 5.12: Venta de refrigeradores trimestral.

La venta parece mostrar un comportamiento estacional, para verificarlo se plantea un modelo en el cual la venta de refrigeradores (FRIG) está

explicada por los trimestres, como se muestra en la [ecuación \(5.14\)](#):

$$\text{FRIG}_t = \alpha_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4 + U_1 t \quad (5.14)$$

Donde:

- **FRIG** representa las ventas de refrigeradores (en miles),
- D_2 es la variable dummy que toma el valor de 1 para los segundos trimestres,
- D_3 es la variable que toma el valor de 1 para los terceros trimestres, y
- D_4 variable que toma el valor de 1 para los cuartos trimestres.

La [Figura 5.13](#) muestra la presencia de efectos estacionales, ya que los coeficientes de las variables D_2 y D_3 son significativos, pues su valor p es inferior al 5%, que es el nivel de significancia considerado.

. regress FRIG i.Dummy						
Source	SS	df	MS	Number of obs = 32		
Model	915635.844	3	305211.948	F(3, 28) = 10.60		
Residual	806142.375	28	28790.7991	Prob > F = 0.0001		
Total	1721778.22	31	55541.2329	R-squared = 0.5318		
				Adj R-squared = 0.4816		
				Root MSE = 169.68		
FRIG	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Dummy						
2	245.375	84.83926	2.89	0.007	71.58966	419.1603
3	347.625	84.83926	4.10	0.000	173.8397	521.4103
4	-62.125	84.83926	-0.73	0.470	-235.9103	111.6603
_cons	1222.125	59.99041	20.37	0.000	1099.24	1345.01

Figura 5.13: Efecto estacional.

La [Figura 5.14](#) presenta la serie original **FRIG** con el efecto estacional y la variable **frides**, la cual representa la variable FRIG sin efecto estacional. La serie desestacionalizada se obtiene sumando a los residuos de la regresión de FRIG en función de las dummy la media aritmética de la variable dependiente FRIG.

La ([Figura 5.14](#)) se logra a través del comando:

Comando

tsline FRIG frides

Finalmente, la incorporación de variables dummy en un modelo, cuando la variable dependiente y algunas explicatorias presentan efecto estacional, es muy importante, pues las dummy eliminan simultáneamente el efecto estacional en dichas variables.

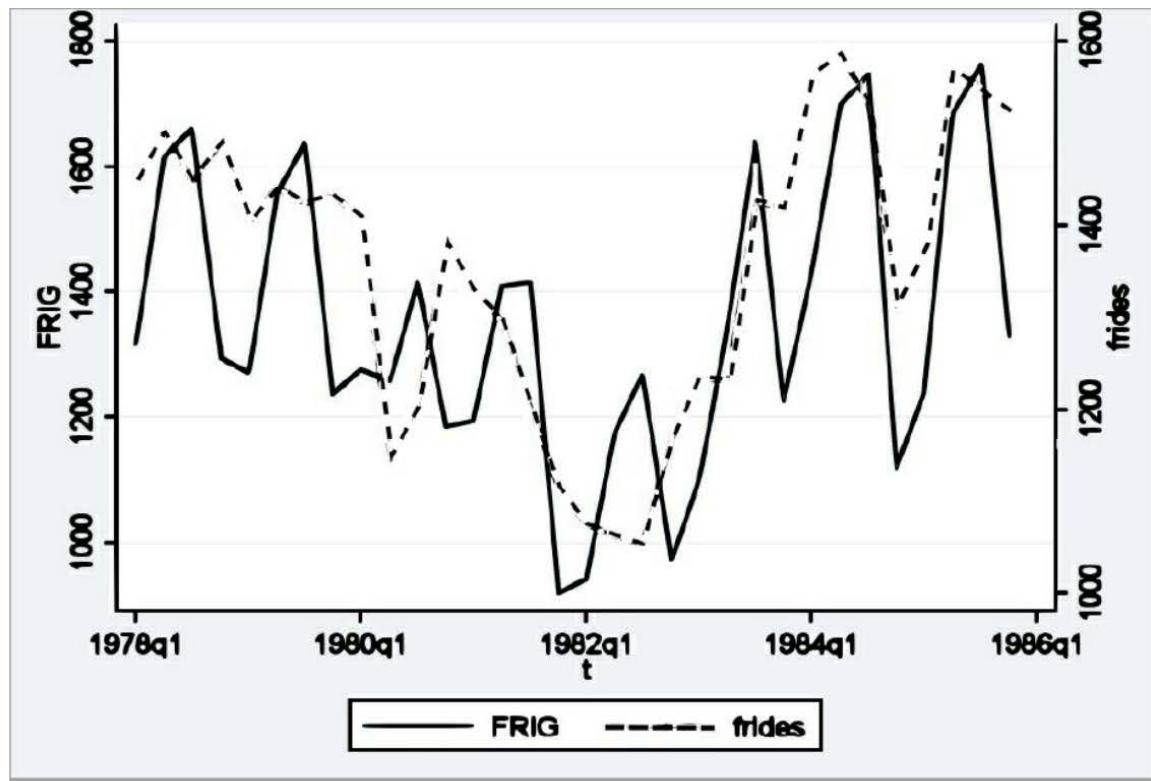


Figura 5.14: Serie original FRIG y desestacionalizada frides.

5.6 Raíz unitaria y cointegración univariada

El análisis de regresión con series de tiempo es de alta complejidad y debe ser tratado en cursos avanzados, no obstante, en este manual se presentan algunos conceptos básicos que ayudan al estudiante o investigador a tener una visión general y a abordar parte de la literatura que hace uso de técnicas y términos especializados sin pretender profundizar en este tema, dado el nivel de formación al cual está enfocado este manual. Un excelente texto que puede servir de referencia para los interesados en documentarse tanto en el tema de regresión con series de tiempo como en el test de raíces unitarias y cointegración es (Enders, 2015).

Cuando se estima un modelo de regresión donde las variables consideradas conforman series de tiempo, se asume implícitamente que dichas series son estacionarias en sentido débil, o sea, su media y su varianza son constantes en el tiempo y su covarianza depende de la longitud del rezago, es decir, $E(Y_t) = \mu$, para todo t , $V(Y_t) = \sigma^2$, para todo t , $\text{Cov}((Y_t - \mu)(Y_{t-k} - \mu)) = \gamma_k$ para todo k . Este supuesto es fundamental en teoría de regresión con series de tiempo, pues su no cumplimiento puede provocar que se generen resultados espurios; por lo tanto, antes de estimar el modelo de regresión es indispensable validar el supuesto de estacionariedad de las series incorporadas en el modelo.

La estacionariedad de una serie de tiempo se puede determinar inicialmente a partir del análisis gráfico de la serie, del análisis de su correlograma o a través de pruebas de raíces unitarias formales, como el test de Dickey-Fuller, Dickey-Fuller aumentado, Philip Perrón o KPSS. Si una serie presenta tendencia y esta parece ser no determinística, se puede sospechar que la serie no es estacionaria; si el correlograma de la función de autocorrelación simple no cae rápidamente a 0 en forma exponencial o sinusoidal después del segundo o tercer rezago, se puede sospechar que la serie no es estacionaria. Aunque existen diversos tests para probar la existencia de raíces unitarias, en este manual solo se considera en detalle el test de Dickey-Fuller y el Dickey-Fuller aumentado.

El test de Dickey Fuller considera tres versiones de una caminata aleatoria: sin deriva y sin tendencia, con deriva y sin tendencia, y con deriva y con tendencia, con hipótesis nula $H_0 : \rho = 1$ versus $H_1 : \rho < 1$, como se muestra en las [ecuaciones \(5.15\)](#) y [\(5.17\)](#).

$$Y_t = \rho Y_{t-1} + \epsilon_t \quad (5.15)$$

$$Y_t = \beta_1 + \rho Y_{t-1} + \epsilon_t \quad (5.16)$$

$$Y_t = \beta_1 + \rho Y_{t-1} + \beta_3 \times \text{Trend}_t + \epsilon_t \quad (5.17)$$

La diferencia entre las tres regresiones está asociada a la presencia de componentes determinísticas en el modelo. En la primera, el modelo representa una caminata aleatoria, en la segunda aparece un término determinístico en el modelo β_1 , y en la tercera aparecen dos términos determinísticos, β_1 y $\beta_3 \times \text{Trend}_t$.

Para mantener la estructura de las hipótesis de no significancia arrojadas por los paquetes estadísticos a las expresiones anteriores se les

resta Y_{t-1} , al factorizar se obtienen las [ecuaciones \(5.18\)](#), [\(5.19\)](#) y [\(5.20\)](#):

$$\Delta Y_t = \delta Y_{t-1} + \epsilon_t \quad (5.18)$$

$$\Delta Y_t = \beta_1 + \delta Y_{t-1} + \epsilon_t \quad (5.19)$$

$$\Delta Y_t = \beta_1 + \delta Y_{t-1} + \beta_3 \text{Trend}_t + \epsilon_t \quad (5.20)$$

Donde:

- $\delta = (\rho - 1)$,
- $\Delta Y_t = Y_t - Y_{t-1}$, y
- ϵ_t es un proceso ruido blanco.

Las hipótesis en esta prueba son $H_0 : \delta = 0$ versus $H_1 : \delta < 0$. Si no se rechaza la hipótesis nula, entonces la serie posee raíz unitaria, es decir, no es estacionaria. Si se rechaza, la serie es estacionaria. En el test de Dickey-Fuller, los estadísticos de prueba ya no se distribuyen bajo H_0 como una t, por eso para tomar la decisión se deben utilizar tablas especiales cuyos valores críticos ya han sido tabulados por Dickey-Fuller. Es importante aclarar que el test de Dickey-Fuller tiene poca potencia cuando el coeficiente que acompaña a la variable dependiente rezagada está cerca a la unidad; igualmente tiene poca potencia cuando no se selecciona el proceso generador correctamente, es decir, no se incluyen las componentes determinísticas adecuadas (Enders, 2015).

Cuando una serie no es estacionaria, se puede volver estacionaria diferenciándola. Si al diferenciarla una vez se vuelve estacionaria, se dice que la serie es integrada de orden 1, **I(1)**; si se debe diferenciar dos veces será **I(2)**, y así sucesivamente. En economía, por lo general, la series de variables nominales son **I(2)**, mientras las reales son **I(1)**. Un aspecto conceptual muy importante, es que ϵ_t se asume que es un proceso ruido blanco, de no serlo, se deben incorporar términos hasta alcanzar ruido blanco, apareciendo el Dickey-Fuller aumentado **DFA**.

Cuando las series que participan en una regresión son no estacionarias, una alternativa es diferenciarlas, estimando el modelo en diferencias. Esta solución elimina las relaciones de largo plazo, por tanto, antes de proceder de esta manera, se debe probar si las series están cointegradas, es decir, si existe una relación de largo plazo, de ser este el caso, la regresión se puede estimar con las series a nivel, debido a que la cointegración elimina las tendencias comunes de las series generando estimaciones superconsistentes. Una prueba de cointegración univariada, sugerida por (Engle y Granger, 1987), es realizar la estimación del modelo de regresión que involucra series **I(1)** y validar si los residuos de dicha regresión son estacionarios, en caso de serlo, las series están cointegradas y la regresión a nivel es válida. Actualmente para probar cointegración se utiliza el test de Johansen y Jolanius (Johansen y Juselius, 1990).

5.6.1 Pruebas de raíz unitaria y cointegración univariada

La teoría y aplicación de los tests de raíz unitaria y cointegración cuando el modelo trabaja con información de series de tiempo, se ha convertido en un paso inevitable en el proceso de modelización, pues dichos tests ayudan a determinar, por un lado, si la regresión debe realizarse en diferencias o a nivel, y si los resultados obtenidos no corresponden a regresiones espurias. En esta sección se presentan algunos elementos introductorios a estos temas, pues un tratamiento detallado de los mismos puede abarcar buena parte de un texto de econometría, además, estos temas se abordan en cursos avanzados de maestría y doctorado.

Raíz unitaria (test de Dickey-Fuller)

Para ilustrar los tests de Dickey-Fuller y Dickey-Fuller aumentado, se considera solo la variación de precios de los Estados Unidos, este mismo proceso debe realizarse a las demás variables incorporadas en el modelo económico. Primero, se sugiere ver la gráfica para sugerir cuál de las tres versiones del test de Dickey-Fuller utilizar, aunque (Enders, 2015) sugiere seguir un algoritmo para determinar la versión a emplear. ([Figura 5.15](#).)

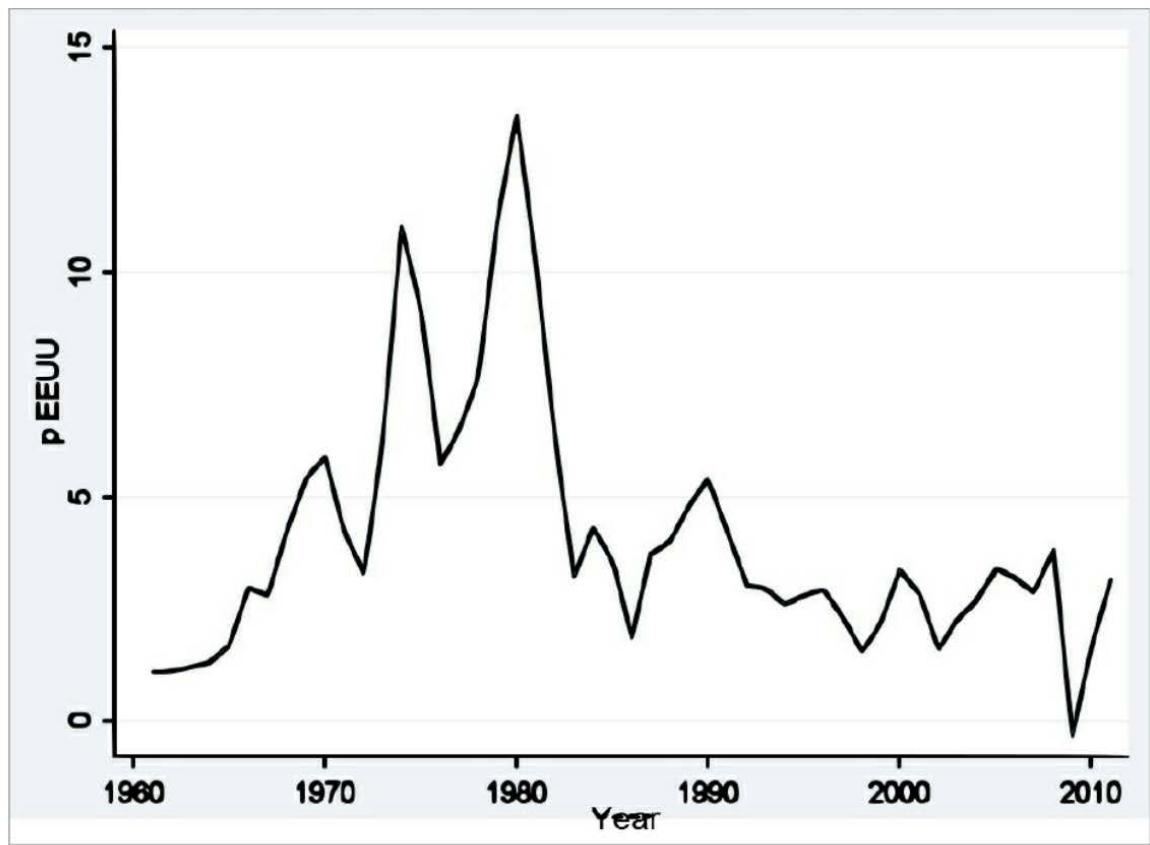


Figura 5.15: Gráfica de la variación de precios de EEUU.

La [Figura 5.15](#) no presenta una tendencia clara, por tanto, escogeremos el modelo de Dickey-Fuller con intercepto pero sin tendencia; adicionalmente, incorporaremos tres términos al modelo para garantizar ruido blanco en el error, es decir, se trabajará con el Dickey-Fuller aumentado. ([Figura 5.16](#).)

dfuller pEEUU, drift regress lags(3)

Augmented Dickey-Fuller test for unit root		Number of obs = 47			
Test Statistic	Z(t) has t-distribution				10% Critical Value
	1% Critical Value	5% Critical Value	10% Critical Value		
Z(t)	-1.839	-2.418	-1.682	-1.302	

-value for Z(t) = 0.0365

D.pEEUU	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pEEUU					
L1.	-.1869138	.1016309	-1.84	0.073	-.3920131 .0181856
LD.	.2728919	.1565224	1.74	0.089	-.0429831 .5887669
L2D.	-.2600664	.1528387	-1.70	0.096	-.5685074 .0483746
L3D.	-.113482	.1658471	-0.68	0.498	-.448175 .2212109
_cons	.8396519	.4976094	1.69	0.099	-.1645645 1.843868

Figura 5.16: Dickey-Fuller aumentado.

Los resultados muestran un valor $p <$ a un nivel de significancia del 5%, por tanto, se rechaza la hipótesis nula y se concluye que la serie es estacionaria.

La ruta a seguir para realizar pruebas de raíces unitarias es:

Ventana

1. Statistics → Time series → Tests → Augmented Dickey-Fuller unit-root test.
2. Elejir la variable que se le quiere realizar la prueba de raíz unitaria, la versión de Dickey-Fuller a utilizar y el número de rezagos a incorporar en el modelo, en caso de que se sospeche que el error no es ruido blanco, en nuestro caso adicionamos tres rezagos.

Cointegración univariada

El concepto de cointegración surge cuando variables integradas del mismo orden generan una combinación lineal integrada de orden menor. En economía, por lo general, las series reales son integradas de orden 1, por lo tanto, dichas series estarán cointegradas si una combinación lineal de ellas es integrada de orden 0, es decir, es estacionaria.

(Engle y Granger, 1987) desarrollan una prueba de cointegración univariada, la cual consiste en evaluar si los residuos generados por un modelo de regresión que incorpora series de variables integradas de orden 1 es integrado de orden 0, es decir, estacionario. La hipótesis nula en esta prueba es que las series de las variables no están cointegradas, mientras la alterna es que están cointegradas.

Probar la existencia de cointegración entre las series de las variables incorporadas en el modelo es importante, porque eso implica la existencia de relaciones de equilibrio de largo plazo y que la regresión se pueda llevar a cabo con las series de las variables a nivel.

Para la cointegración univariada planteada por (Engle y Granger, 1987), se sigue la ruta explicada para probar existencia de raíces unitarias, en esta ocasión sobre los residuos del modelo estimado, teniendo cuidado de no utilizar los valores críticos de significancia de Dickey-Fuller o Dickey-Fuller aumentado, ya que no son apropiados; por ello, Engle y Granger calcularon los valores críticos asintóticos, los cuales son -3.34 y -3.04 para niveles de significancia del 5% y 10%, respectivamente.

```

. predict resid, residuals

. tset Year, yearly
      time variable: Year, 1961 to 2011
                  delta: 1 year

. dfuller resid, drift lags(1)

Augmented Dickey-Fuller test for unit root           Number of obs = 49
                                                       Z(t) has t-distribution
Test          1% Critical    5% Critical    10% Critical
Statistic     Value         Value         Value
Z(t)          -3.721       -2.410       -1.679      -1.300
p-value for Z(t) = 0.0003

```

Figura 5.17: Prueba de cointegración de Granger.

Como se observa en la [Figura 5.17](#), los residuos asociados a la regresión mostrada en la [Figura 5.1](#) rechazan la hipótesis nula, por tanto, las series de las variables están cointegradas y se puede correr el modelo a nivel con sus interpretaciones correspondientes.

Ecuaciones simultáneas

Temas Tratados

6.1 Modelos VAR

Una pregunta que surge en este capítulo es: ¿Cómo se trabajan los modelos econométricos cuando las variables se causan mutuamente, es decir, cuando X influye en Y , pero a su vez Y influye en X ? La respuesta es: diseñando modelos de ecuaciones simultáneas, los cuales consideran esa mutua causalidad.

En este capítulo se abordan los modelos de ecuaciones simultáneas siguiendo el texto de (Gujarati y Porter, 2010). Ahora bien, si se parte del hecho de que las variables económicas interactúan frecuentemente entre sí por formar parte de un sistema económico, los modelos uniecuacionales considerados hasta el momento en este manual están limitados porque permiten estudiar la dirección de causalidad en un solo sentido, de las variables explicatorias X hacia Y , pero no de la variable dependiente Y hacia algunas explicatorias X . Esta situación plantea la necesidad de diseñar nuevos modelos que consideren dicha interacción en doble sentido, apareciendo así las ecuaciones simultáneas. Lo anterior significa que van a aparecer variables aleatorias como explicatorias y, por lo tanto, no pueden estar correlacionadas con el error de la ecuación correspondiente. Sin embargo, se demuestra que en un sistema de ecuaciones simultáneas puro las variables explicatorias aleatorias están correlacionadas con el error, lo que implica que los **MCO** no pueden ser utilizados porque los estimadores obtenidos son sesgados e inconsistentes. Lo anterior se resume en el siguiente modelo de ecuaciones simultáneas conformado por las [ecuaciones \(6.1\)](#) y [\(6.2\)](#).

$$Y_{1t} : \beta_1 + \beta_2 Y_{2t} + \beta_3 X_{1t} + U_{1t} \quad (6.1)$$

$$Y_{2t} : \lambda_1 + \lambda_2 Y_{1t} + \lambda_3 X_{2t} + U_{2t} \quad (6.2)$$

Como se observa, Y_{1t} y Y_{2t} son variables explicatorias aleatorias, pero a su vez son variables explicadas; intuitivamente se puede ver que existe autocorrelación entre Y_{2t} y U_{1t} , así como entre Y_{1t} y U_{2t} , pues si se mueve U_{1t} , se mueve Y_{1t} y si se mueve Y_{1t} , se mueve Y_{2t} , pues esta depende de Y_{1t} ; es decir, un movimiento en U_{1t} desencadena un movimiento en Y_{2t} , lo que indica que la covarianza $(Y_{2t}, U_{1t}) \neq 0$, igual sucede con movimientos en U_{2t} , generando con ello que la covarianza $(Y_{1t}, U_{2t}) \neq 0$.

La correlación que se presenta en ecuaciones simultáneas entre las variables explicatorias aleatorias y el error, como ya se comentó, generan que los estimadores **MCO** sean sesgados e inconsistentes.

En un sistema de ecuaciones simultáneas intervienen dos tipos de variables: endógenas, que se determinan dentro del sistema, y predeterminadas, que ya vienen dadas y están conformadas por las exógenas, las exógenas rezagadas y las endógenas rezagadas.

Para considerar una endógena rezagada como no aleatoria se asume que no existe autocorrelación en el error de la ecuación en la cual se encuentra la endógena rezagada; de no ser el caso, dicha variable dependiente rezagada no puede considerarse predeterminada.

Las variables endógenas son aleatorias y con frecuencia se representan con Y , mientras la predeterminadas son no aleatorias y con frecuencia se denotan con X .

A continuación se presenta un sistema de ecuaciones simultáneas hipotético conformado por las [ecuaciones \(6.3\)](#), [\(6.4\)](#) y [\(6.5\)](#):

$$Y_{1t} = \beta_{10} + \beta_{12} Y_{2t} + \gamma_{11} X_{1t} + \gamma_{12} X_{2t} + \gamma_{13} X_{3t} + U_{1t} \quad (6.3)$$

$$Y_{2t} = \beta_{20} + \beta_{21} Y_{1t} + \beta_{22} Y_{3t} + \gamma_{21} X_{3t} + \gamma_{22} X_{4t} + U_{2t} \quad (6.4)$$

$$Y_{3t} = \beta_{30} + \beta_{32} Y_{2t} + \gamma_{31} X_{1t} + \gamma_{32} X_{2t} + \gamma_{33} Y_{3t-1} + U_{3t} \quad (6.5)$$

Las ecuaciones que conforman el sistema anterior se denominan ecuaciones estructurales porque representan la estructura o comportamiento de un sistema. En dicho sistema, las variables endógenas son Y_{1t} , Y_{2t} y Y_{3t} , mientras las predeterminadas son X_{1t} , X_{2t} , X_{3t} , X_{4t} y Y_{3t-1} . Un análisis sencillo a partir de los movimientos en las perturbaciones de cada ecuación muestra la existencia de correlación entre las perturbaciones y las variables endógenas que aparecen en este modelo como explicatorias, imposibilitando el uso de **MCO**. Una alternativa de estimación es utilizar las denominadas ecuaciones reducidas, que consisten en expresar las variables endógenas en función de las variables predeterminadas y del error. Existirán tantas ecuaciones reducidas como variables endógenas haya en el sistema; por ejemplo, para el sistema anterior, la ecuación reducida para Y_{1t} es la [ecuación \(6.6\)](#):

$$Y_{1t} = \pi_0 + \pi_1 X_{1t} + \pi_2 X_{2t} + \pi_3 X_{3t} + \pi_4 X_{4t} + \pi_5 Y_{3t-1} + W_{1t} \quad (6.6)$$

Las ecuaciones reducidas pueden ser estimadas por **MCO**, dado el supuesto de que las predeterminadas no sean aleatorias. Una estrategia cuando las ecuaciones son exactamente identificadas, es estimar por **MCO** los coeficientes reducidos y a partir de ellos estimar los estructurales. Este procedimiento se denomina mínimos cuadrados indirectos (**MCI**), desafortunadamente este método proporciona las estimaciones puntuales de los coeficientes estructurales, pero no sus errores estándar. Sin embargo, en caso de requerirlos, los errores estándar se pueden hallar mediante un procedimiento que presenta alta dificultad, lo que a su vez genera que realizar las inferencias también sea difícil. Un método alternativo que produce las mismas estimaciones, pero con sus errores estándar, es el método de mínimo cuadrados ordinarios en dos etapas (**MC2E**). Este método se denomina así porque en una primera etapa estima por **MCO** la forma reducida asociada a la variable endógena que aparece como explicatoria, y en una segunda etapa vuelve a estimar por **MCO** la ecuación de interés, pero reemplazando los valores de la variable endógena por sus valores predichos obtenidos a partir de la estimación de su forma reducida. Este procedimiento produce estimaciones consistentes de los parámetros; en este sentido, el **MC2E** se constituye en un método de estimación por variables instrumentales y funciona bien en la medida que el tamaño de la muestra sea grande, en muestras pequeñas los resultados obtenidos deben tomarse con precaución.

Caso de estudio. Para ilustrar el método de estimación en dos etapas (**MC2E**), considérese el siguiente sistema de ecuaciones conformado por las [ecuaciones \(6.7\)](#) y [\(6.8\)](#):

$$VTCN_t = \beta_0 + \beta_1 \times PCOL_t + \beta_2 \times CRECOL_t + U_{1t} \quad (6.7)$$

$$PCOL_t = \alpha_0 + \alpha_1 \times VTCN_t + \alpha_3 \times PEEUU_t + U_{2t} \quad (6.8)$$

Donde:

- número de ecuaciones: 2;
- variable endógena **VTCN**: Variación de tasa de cambio en Colombia;
- variable endógena **PCOL**: inflación colombiana;
- variable predeterminada **CRECOL**: crecimiento en Colombia;
- variable predeterminada **PEEUU**: inflación en Estados Unidos.

Por medio de las condiciones de orden concluimos que ambas ecuaciones son exactamente identificadas porque cada una de ellas excluye una variable, es decir, la primera ecuación excluye a **PEEUU** y la segunda excluye a **CRECOL**.

La estimación por mínimos cuadrados en dos etapas de las [ecuaciones \(6.7\)](#) y [\(6.8\)](#) del sistema anterior, usando Stata, se muestra en las siguientes instrucciones:

Comando

```
ivregress 2sls VTCN (PCOL = CRECOL PEEUU) CRECOL
```

Comando

```
ivregress 2sls PCOI (VTCN = CRECOL PEEUU) PEEUU
```

Como se observa, para construir la variable instrumental se escribe la variable endógena de la ecuación y entre paréntesis la variable endógena explicatoria igualada a las variables predeterminadas del sistema, seguidamente, se introduce el resto de variables explicatorias de la ecuación, si existen. Las [Figuras 6.1](#) y [6.2](#) muestran los resultados obtenidos para la primera ecuación y la segunda ecuación del sistema respectivamente, en ellos se observa que en la primera ecuación, el crecimiento económico de Colombia no es un determinante para explicar la variación del tipo de cambio, mientras que la inflación en Colombia sí lo es. Para la segunda ecuación se observa que la inflación de Estados Unidos y la variación del tipo de cambio son determinantes de la inflación en Colombia, esto porque los valores p asociados son inferiores a un nivel de significancia del 5%.

```
. ivregress 2sls VTCN ( pCOL= pEEUU pEEUU CreCol) CreCol
```

note: pEEUU omitted because of collinearity

Instrumental variables (2SLS) regression

Number of obs	=	51
Wald chi2(2)	=	12.97
Prob > chi2	=	0.0015
R-squared	=	0.3457
Root MSE	=	9.5219

VTCN	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
pCOL	.2955046	.3045844	0.97	0.332	-.3014699 .8924791
CreCol	-1.998455	.6406969	-3.12	0.002	-3.254198 -.7427122
_cons	16.27667	6.343081	2.57	0.010	3.844459 28.70888

Instrumented: pCOL

Instruments: CreCol pEEUU

Figura 6.1: Estimación por mínimos cuadrados en dos etapas para la variable VCTN.

```
. ivregress 2sls pCOL ( VTCN = pEEUU CreCol) pEEUU
```

Instrumental variables (2SLS) regression

Number of obs	=	51
Wald chi2(2)	=	29.21
Prob > chi2	=	0.0000
R-squared	=	0.5382
Root MSE	=	5.9713

pCOL	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
VTCN	.3492357	.1760201	1.98	0.047	.0042427 .6942287
pEEUU	1.406904	.3006005	4.68	0.000	.817738 1.99607
_cons	6.041059	2.50381	2.41	0.016	1.133682 10.94844

Instrumented: VTCN

Instruments: pEEUU CreCol

Figura 6.2: Estimación por mínimos cuadrados en dos etapas para la variable PCOLS.

Por lo tanto, las estimaciones para las dos ecuaciones del sistema planteado vienen dadas por las siguientes ecuaciones:

$$\widehat{\text{VTCN}}_t = 16,27 + 0,2955 \times \text{PCOL}_t - 1,998 \times \text{CRECOL}_t \quad (6.9)$$

$$\widehat{\text{PCOL}}_t = 6,04 + 0,3492 \times \text{VTCN}_t + 1,45 \times \text{PEEUU}_t \quad (6.10)$$

La interpretación de dichas estimaciones se realiza de la misma manera como se ha ilustrado en capítulos anteriores.

6.1

Modelos VAR

Los modelos de vectores autorregresivos conocidos como **VAR**, pueden entenderse como una generalización de los modelos autorregresivos univariados, pues en ellos hay más de una variable dependiente donde cada una de ellas es explicada por sus valores rezagados y por los valores rezagados de todas las demás variables dependientes. La estructura de un modelo **VAR** bivariado de orden k se presenta en las [ecuaciones \(6.11\)](#) y [\(6.12\)](#).

$$Y_{1t} = \beta_{10} + \beta_{11}Y_{1t-1} + \cdots + \beta_{1k}Y_{1t-k} + \alpha_{11}Y_{2t-1} + \cdots + \alpha_{1k}Y_{2t-k} + U_{1t} \quad (6.11)$$

$$Y_{2t} = \beta_{20} + \beta_{21}Y_{2t-1} + \cdots + \beta_{2k}Y_{2t-k} + \alpha_{21}Y_{1t-1} + \cdots + \alpha_{2k}Y_{1t-k} + U_{2t} \quad (6.12)$$

Donde, u_{it} es ruido blanco. La flexibilidad y fácil generalización de los modelos **VAR** permite abarcar en su estructura modelos de media móvil, dando origen en la versión multivariada a los **VARMA**.

Un ejemplo de un **VAR** bivariado donde solo hay dos variables endógenas o dependientes Y_{1t} y Y_{2t} , y un solo rezago se muestra en las [ecuaciones \(6.13\)](#) y [\(6.14\)](#).

$$Y_{1t} = \beta_{10} + \beta_{11}Y_{1t-1} + \alpha_{11}Y_{2t-1} + U_{1t} \quad (6.13)$$

$$Y_{2t} = \beta_{20} + \beta_{21}Y_{2t-1} + \alpha_{21}Y_{1t-1} + U_{2t} \quad (6.14)$$

En los modelos **VAR**, los términos de error o perturbaciones se denominan choques o impulsos. En cuanto a la estimación, los modelos **VAR** se pueden considerar como un sistema de ecuaciones simultáneas donde las variables explicatorias son predeterminadas o no estocásticas, y por ende no están correlacionadas con el error, además, si el error no presenta autocorrelación, los **MCO** pueden ser utilizados como método de estimación de los modelos **VAR**.

Por otro lado, los modelos **VAR** que se tratan frecuentemente en cursos avanzados de series de tiempo múltiples, se diferencian de las ecuaciones simultáneas, en que todas las variables se consideran endógenas, adicionalmente, los pronósticos con un modelo **VAR** son más precisos que los realizados con los sistemas de ecuaciones tradicionales.

Finalmente, entre las ventajas de los modelos **VAR** se encuentran: el uso de las funciones impulso-respuesta que permiten estudiar los efectos de un choque sobre las variables del modelo **VAR**, y la descomposición de varianza que ayuda a determinar cuánto de los movimientos de una variable, por ejemplo y_1 , afecta la varianza de otra variable, por ejemplo y_2 . Sin embargo, los modelos **VAR** han sido cuestionados porque son ateóricos, como los modelos **ARMA**, y, además, no se tienen criterios claros sobre la longitud de los rezagos a incluir en el modelo, aunque esto se soluciona haciendo uso de los criterios de información como **Akaike** o **Schwartz**. Otra desventaja de los modelos **VAR** es que consume demasiados grados de libertad, en tanto el número de variables involucradas sea alto, así como el número de rezagos de cada variable considerada. Los **VAR** suponen que todas las variables involucradas son estacionarias; de no serlo, deben llevarse a estacionarias con todas sus implicaciones teóricas.

Para ilustrar los modelos **VAR**, sin pretender profundizar en ellos, se han escogido las variables variación de tipo de cambio (**VTCN**) e inflación de Colombia (**pCOL**), las cuales se rezagaron dos veces, para la estimación de este modelo en Stata se debe seguir esta ruta por ventanas:

Ventana

1. Statistics → multivariate time series → vector autoregression → 
2. Definir las variables dependientes que en nuestro caso serán **VTCN** y **pCOL**
3. Colocar el número de rezagos y las variables exógenas si existen en el modelo (Al incluir este proceso el **VAR** se conoce como **VARX**).

La [Figura 6.3](#) muestra el resultado obtenido en Stata.

Sample: 1963 - 2011		No. of obs	=	49	
Log likelihood = -322.3292		AIC	=	13.56446	
FPE = 2670.786		HQIC	=	13.71094	
Det(Sigma_m1) = 1773.196		SBIC	=	13.95054	
Equation	Parms	RMSE	R-sq	chi2	P>chi2
VTCN	5	9.256	0.4464	39.50789	0.0000
pCOL	5	5.60046	0.6306	83.66386	0.0000
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
VTCN					
VTCN					
L1.	.4612627	.1545644	2.98	0.003	.1583221 .7642034
L2.	.0913698	.1569343	0.58	0.560	-.2162157 .3989554
pCOL					
L1.	-.1363946	.2554657	-0.53	0.593	-.6370982 .3643091
L2.	.4592169	.2589472	1.77	0.076	-.0483103 .9667441
_cons	.3590934	2.879115	0.12	0.901	-5.283868 6.002055
pCOL					
VTCN					
L1.	-.0564781	.0935211	-0.60	0.546	-.2397761 .12682
L2.	.0378455	.0949551	0.40	0.690	-.148263 .223954
pCOL					
L1.	.5819904	.1545728	3.77	0.000	.2790334 .8849475
L2.	.285144	.1566793	1.82	0.069	-.0219418 .5922297
_cons	2.375707	1.742045	1.36	0.173	-1.038638 5.790052

Figura 6.3: Estimación de un VAR(2) bivariado.

Vale la pena destacar que las estimaciones y las gráficas de las funciones impulsorespuesta, tema no tratado en este manual, se obtienen siguiendo la ruta:



Microeconometría

Temas Tratados

7.1 Modelos de elección discreta binaria

- 7.1.1 Modelo lineal de probabilidad
- 7.1.2 El modelo logit
- 7.1.3 El modelo probit
- 7.1.4 Modelos de datos de recuento
- 7.1.5 Modelo de regresión de Poisson
- 7.1.6 El modelo de regresión Tobit

La pregunta en este capítulo es: ¿Cómo se modela cuando la variable dependiente es cualitativa? La respuesta es: si bien en todos los modelos planteados hasta el momento la variable dependiente es cuantitativa, en ocasiones surgen modelos en los cuales la variable de respuesta es cualitativa con dos o más categorías, mientras que las variables explicatorias pueden ser cualitativas y/o cuantitativas. Estos modelos forman parte de los denominados modelos de variable dependiente limitada, porque el número de valores que puede tomar la variable dependiente es limitado; así por ejemplo, cuando la variable dependiente Y es cualitativa con dos categorías, los valores que puede tomar Y en escala nominal son solo dos, y cuando la variable dependiente Y es cualitativa con cinco categorías, los valores que puede tomar Y en escala nominal u ordinal son solo cinco. Un ejemplo de orden práctico es: ¿Cuáles son los determinantes que explican la deserción estudiantil? En este caso, la variable dependiente $Y = \text{deserción}$ toma dos categorías, 1 si es desertor y 0 si no es un desertor, tales modelos se conocen como variable dependiente binaria o dicótoma. Si la variable dependiente tiene cuatro categorías, un ejemplo sería: ¿cuáles son los determinantes que explican la calificación crediticia de los clientes de una institución financiera? En este ejemplo, la variable dependiente $Y = \text{calificación crediticia}$ puede tomar más de dos categorías, 1 si es mal cliente, 2 si es un cliente aceptable, 3 si es un cliente bueno, y 4 si es un cliente excelente, este caso se conoce como variable de respuesta policótoma o de categorías múltiples.

Existen otros modelos de variable dependiente limitada, como son los modelos de conteo o de Poisson, en los cuales los valores que toma la variable dependiente son discretos y limitados, como por ejemplo, cuando se desea encontrar los factores que explican el número de caídas de un anciano, o los determinantes del número de libros leídos en un año. Otros modelos de variable dependiente limitada son los modelos Tobit, los cuales se utilizan cuando una gran parte de la muestra seleccionada presenta el valor de 0 en la variable dependiente, pero la muestra restante, para la cual sí se tiene la información de la variable dependiente Y , sigue una distribución normal. En estos modelos se dice que la variable dependiente presenta una solución de esquina. Los modelos de regresión truncada y censurada también forman parte de los modelos de variable dependiente limitada.

Cuando los datos no están agrupados, la estimación de un modelo logit y probit debe realizarse por el método de máxima verosimilitud, dado que dichos modelos no son lineales en sus parámetros. Este método genera estimadores consistentes, asintóticamente normales y asintóticamente eficientes. Por otro lado, para realizar las pruebas de hipótesis alrededor de los coeficientes de regresión de un modelo logit y probit se utilizan métodos que consideran tanto la linealidad como la no linealidad de los parámetros del modelo, dichos métodos son: la prueba de razón verosimilitud, de Wald y del multiplicador de Lagrange. Estas pruebas en grandes muestras conducen a resultados similares y cada una de ellas sigue una distribución **chi-cuadrado**.

Los modelos de elección discreta pueden interpretarse desde dos puntos de vista: el primero es a través de efectos marginales, y el segundo es a través de las razones **odds**, estas últimas juegan un papel importante cuando el objetivo es la comparación entre distintas situaciones. En este manual, para los modelos presentados, se analizará solo el efecto marginal, entendido como los cambios en probabilidad de que suceda el evento de interés ante un cambio en la variable explicatoria asociada.

7.1

Modelos de elección discreta binaria

En esta sección se considera solo el caso donde la variable de respuesta es binaria. Una característica de los modelos donde la variable dependiente es cualitativa, es que el valor medio o esperado de la variable dependiente Y condicionado al conjunto de variables explicatorias produce una probabilidad, de allí que dichos modelos también sean conocidos como modelos de probabilidad.

7.1.1 Modelo lineal de probabilidad

La estructura de este modelo es similar a los modelos con variable dependiente cuantitativa, la única diferencia es que la variable dependiente en este caso es cualitativa y, por ende, el valor esperado de la variable dependiente **Y** condicionado a un conjunto de variables explicatorias **X** proporciona una probabilidad. Aunque este modelo tiene una serie de limitaciones, como la probabilidad, que en ocasiones no cae entre 0 y 1, los errores no siguen una distribución normal sino una Bernoulli, la varianza del error es heterocedástica y la probabilidad es lineal a **X**, es decir, el efecto que tiene **X** sobre la probabilidad de que se presente el evento de interés es constante, independiente del valor que tome la variable **X**. Esta última limitación plantea la necesidad de desarrollar modelos que no se relacionen linealmente con **X**, apareciendo así los modelos logit y los probit, los cuales utilizan las funciones de distribución acumulada asociadas a variables aleatorias con función de densidad logística y normal, respectivamente.

Caso de estudio. Suponga que se desea identificar los determinantes que una persona sea profesional o no, en este caso, la variable dependiente es binaria o dicótoma, tomando el valor de 1 si la persona es profesional, y 0 si no lo es, como variables explicatorias se consideran: el ingreso familiar; la zona, que toma el valor de 1 si reside en Bogotá y 0 para el resto del país; y el sexo, que toma el valor de 1 si es hombre y 0 si es mujer. Para estimar este modelo por mínimos cuadrados ordinarios se utilizan datos de la encuesta del mercado laboral en Colombia del periodo 1967-1970, elaborada por la universidad de los Andes, procediendo de manera similar a como se estima un modelo lineal estándar, adicionando, después de las variables explicatorias, la opción robust, la cual permite obtener estimadores robustos a la heterocedasticidad. Usando comandos se procede de la siguiente manera:

Comando

```
regress Profesional Ingresofamiliar Sexo Zona, robust
```

La [Figura 7.1](#) muestra el resultado obtenido por Stata para el modelo de probabilidad lineal en el caso de estudio considerado.

```
- regress Profesional Ingresofamiliar Sexo Zona, robust
```

Linear regression		Number of obs		= 55,757	
		F(3, 55753)		= 188.63	
		Prob > F		= 0.0000	
		R-squared		= 0.0153	
		Root MSE		= .145	
		Robust			
Profesional		Coef.	Std. Err.	t	P> t
Ingresofamiliar		7.41e-07	5.90e-08	12.55	0.000
Sexo		.0274826	.0012863	21.37	0.000
Zona		.0157778	.0012301	12.83	0.000
_cons		-.0069079	.0011183	-6.18	0.000

Figura 7.1: Estimación de un modelo de probabilidad lineal.

Los resultados anteriores se asemejan a los producidos por un modelo de regresión cuando la variable explicada es cuantitativa, no obstante, en este caso los valores ajustados de la variable dependiente son probabilidades. Una ventaja que presentan los modelos de probabilidad lineal frente al logit y al probit, es que los coeficientes asociados a las variables explicatorias miden el cambio en probabilidad de que se presente el evento de interés, ante cambios en la variable explicatoria correspondiente. Por ejemplo, un incremento en el ingreso familiar de 1000 pesos, incrementa la probabilidad de ser profesional en 0.0000741%, aunque el coeficiente asociado al ingreso familiar es significativo estadísticamente, desde el punto de vista práctico es irrelevante. Para un nivel de ingreso familiar dado, residir en Bogotá incrementa la probabilidad de ser profesional en 1.57% frente a un no residente en Bogotá, independientemente de si es hombre o mujer, mientras que dado un nivel de ingreso familiar, ser hombre incrementa la probabilidad de ser profesional 2.74% frente a una mujer, independientemente del lugar de residencia.

7.1.2 El modelo logit

El modelo logit aparece para solucionar dos problemas centrales asociados al modelo de probabilidad lineal: el primero, que algunas probabilidades estimadas no se encuentren entre 0 y 1; y el segundo, que el efecto que tiene la variable explicatoria sobre la probabilidad de un evento particular sea constante, independientemente del valor que tome la variable explicatoria. El modelo logit garantiza, por un lado, que las probabilidades estén entre 0 y 1, pues utiliza la distribución de probabilidad acumulada de una variable aleatoria con función de densidad logística, y por el otro, que al utilizar el logit la función de distribución acumulada, la cual tiene forma de s, no permite que el efecto que tiene una variable explicatoria sobre

la probabilidad sea constante, pues en estos modelos hay tramos en los que la probabilidad de que se presente el evento de interés crece de forma muy lenta, principalmente cuando la variable explicatoria toma valores muy bajos o muy altos, mientras existen otros tramos donde la probabilidad de que se presente el evento crece muy rápidamente.

Para ilustrar la estimación de logit no agrupados utilizaremos los datos del modelo lineal de probabilidad trabajados en la subsección anterior, es decir, se trata de identificar los determinantes de ser profesional, tomando como variable dependiente una variable dicótoma que toma el valor de 1 si la persona es profesional, y 0 si no es profesional. Como variables explicatorias se tienen: el ingreso familiar; la zona, que toma el valor de 1 si reside en Bogotá y 0 si reside en otra zona del país; y el sexo, que toma el valor de 1 si es hombre y 0 si es mujer. El modelo planteado se presenta en la [ecuación \(7.1\)](#).

$$P(y = 1/x) = \frac{\exp(\beta_0 + \beta_1 \times \text{ingresofamiliar} + \beta_2 \times \text{zona} + \beta_3 \times \text{sexo})}{(1 + \exp(\beta_0 + \beta_1 \times \text{ingresofamiliar} + \beta_2 \times \text{zona} + \beta_3 \times \text{sexo}))} + U_t \quad (7.1)$$

Para la estimación del logit en general se sigue esta ruta por ventanas:

Ventana

statistics → binary outcomes → logistic regression → definir la variable dependiente e Independientes → Ok.

Por comando sería:

Comando

logit variable dependiente lista de variables independientes

Como sugieren (Cameron y Trivedi, 2005), para estimar el modelo logit no se debe incorporar la opción **robust**, si las observaciones son independientes y $F(X,\beta)$ está bien especificado. Cuando se introduce la opción **robust** en la estimación del modelo logit, y este genera errores estándar muy diferentes a los obtenidos cuando no se introduce la opción **robust** en la estimación, eso significa posibles errores de especificación en el modelo. Si no es posible conseguir que las observaciones sean totalmente independientes, los errores estándar deben ajustarse introduciendo el comando **vce(cluster clustvar)**. La estimación del logit para el caso de estudio por comando será:

Comando

logit Profesional Ingresofamiliar Sexo Zona

```
. logit Profesional Ingresofamiliar Sexo Zona
```

```
Iteration 0:  log likelihood = -5858.1761
Iteration 1:  log likelihood = -5731.687
Iteration 2:  log likelihood = -5467.5697
Iteration 3:  log likelihood = -5463.6615
Iteration 4:  log likelihood = -5463.6552
Iteration 5:  log likelihood = -5463.6552
```

```
Logistic regression                                         Number of obs      =      55757
                                                               LR chi2(3)        =     789.04
                                                               Prob > chi2       =     0.0000
Log likelihood = -5463.6552                                Pseudo R2         =     0.0673
```

Profesional	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Ingresofamiliar	.0000188	1.15e-06	16.28	0.000	.0000165 .000021
Sexo	1.426537	.0694952	20.53	0.000	1.290329 1.562745
Zona	.9756514	.0939643	10.38	0.000	.7914847 1.159818
cons	-5.65372	.1065337	-53.07	0.000	-5.862523 -5.444918

Figura 7.2: Estimación logit.

Los resultados del logit ([Figura 7.2](#)) muestran un efecto positivo sobre la probabilidad de ser profesional, del ingreso familiar, la zona y el sexo, siendo estas variables explicativas significativas en el modelo. El signo de las variables en este modelo coincide con el de los efectos marginales. El log converge rápidamente en solo cinco iteraciones, un número elevado de iteraciones es indicio de posibles problemas de multicolinealidad entre las variables explicativas. Para probar una hipótesis en un modelo logit existen dos procedimientos: el test de Wald y el test de razón de verosimilitud, siendo este último equivalente al test de Wald si el modelo está bien especificado. Los resultados muestran que el modelo estimado es significativo globalmente, pues el valor p del estadístico chi cuadrado es 0.000, inferior a un nivel de significancia del 5%. El modelo logit utiliza una medida de ajuste denominada pseudo R^2 , la cual se interpreta de manera similar al coeficiente de determinación del modelo de regresión lineal ya tratado.

Después de estimar el modelo logit se pueden hallar los efectos marginales, es decir, los cambios en probabilidad ante cambios en las variables explicativas. Stata presenta tres formas para calcular el efecto marginal: el efecto marginal en un valor representativo, el efecto marginal en la media y el efecto marginal promedio. El efecto marginal proporciona mucha más información que los coeficientes de regresión del modelo logit. El efecto marginal en un valor representativo se obtiene después de estimar el modelo logit, y proporciona el efecto marginal para un valor específico del vector de variables explicativas. Cuando no se define un valor específico para el vector de variables explicativas, el Stata calcula el efecto marginal por *default*, tomando como valor para cada variable explicatoria su promedio; este procedimiento se conoce como efecto marginal para el agente promedio. El efecto marginal promedio se calcula tomando el efecto marginal para cada individuo y luego se promedian dichos efectos marginales; este procedimiento se conoce como efecto marginal para el promedio de los agentes. Los comandos para calcular los efectos marginales después de la estimación de logit se describen a continuación:

- `mfx, at(X = X*Z = Z*W = W*)` efecto marginal en los valores representativos X^*yW^* , fijados por el investigador;
- `mfx` efecto marginal calculado en la media de cada variable explicatoria; y
- `margeff` efecto marginal promedio.

Es importante aclarar que el comando `margeff` por lo general no se encuentra instalado en el Stata, por lo cual debe instalarse a través del comando `ya comentado: findit margeff`

Comando

```
mfx, at(Ingresofamiliar=200 Sexo=1 Zona=1)
```

La [Figura 7.3](#) muestra el efecto marginal para un conjunto de valores representativos del caso de estudio.

```
. mfx, at( Ingresofamiliar =200 Sexo=1 Zona=1)
```

Marginal effects after logit

```
y = Pr(Profesional) (predict)  
= .03740678
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	x
Ingres~r	6.76e-07	.00000	16.20	0.000	5.9e-07	7.6e-07		200
Sexo*	.0281612	.00138	20.45	0.000	.025462	.030861		1
Zona*	.0229699	.00171	13.40	0.000	.01961	.026329		1

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Figura 7.3: Efectos marginales.

Los resultados muestran que cuando el ingreso familiar pasa de 200,000 pesos a 201,000 pesos, la probabilidad de ser profesional se incrementa en tan solo 0,0000655%, residir en Bogotá incrementa la probabilidad de ser profesional en 2.29% con relación a un no residente en Bogotá, mientras que ser hombre incrementa la probabilidad en 2.81% de ser profesional con relación a una mujer.

Es importante aclarar que el comando **margins** también permite calcular los efectos marginales(Cameron y Trivedi, 2009), pero para hacerlo adecuadamente en la estimación del modelo logit, las variables binarias deben señalarse anteponiendo la letra **i**, como se muestra a continuación para el caso de estudio:

Comando

```
logit Profesional Ingresofamiliar i.Sexo i.Zona
```

```
. logit Profesional Ingresofamiliar i.Sexo i.Zona
```

```
Iteration 0:  log likelihood = -5858.1761
Iteration 1:  log likelihood = -5731.687
Iteration 2:  log likelihood = -5467.5697
Iteration 3:  log likelihood = -5463.6615
Iteration 4:  log likelihood = -5463.6552
Iteration 5:  log likelihood = -5463.6552
```

```
Logistic regression                                         Number of obs      =  55,757
                                                               LR chi2(3)        =   789.04
                                                               Prob > chi2       = 0.0000
Log likelihood = -5463.6552                                Pseudo R2         =  0.0673
```

Profesional	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Ingresofamiliar	.0000188	1.15e-06	16.28	0.000	.0000165	.000021
Sexo						
Hombre	1.426537	.0694952	20.53	0.000	1.290329	1.562745
Zona						
Bogotá	.9756514	.0939643	10.38	0.000	.7914847	1.159818
_cons	-5.65372	.1065337	-53.07	0.000	-5.862523	-5.444918

Figura 7.4: Estimación logit para el margins.

Como se muestra en la [figura 7.4](#), los resultados obtenidos coinciden con la estimación estándar del logit para el caso de estudio. Para calcular los efectos marginales a través del margins, se utiliza el siguiente comando para el caso de estudio:

Comando

```
margins, dydx(*) at(Ingresofamiliar=200 Sexo=1 Zona=1)
```

Como se muestra en la [Figura 7.5](#), el cálculo de los efectos marginales para un conjunto de valores representativos usando margins, coinciden con los obtenidos a través del comando **mfx**.

```
. margins, dydx(*) at( Ingresofamiliar =200 Sexo=1 Zona=1)

Conditional marginal effects
Number of obs      =      55,757
Model VCE       : OIM

Expression   : Pr(Profesional), predict()
dy/dx w.r.t. : Ingresofamiliar 1. Sexo 1. Zona
at           : Ingresofam-r      =      200
                  Sexo          =      1
                  Zona          =      1
```

	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
Ingresofamiliar	6.76e-07	4.17e-08	16.20	0.000	5.94e-07	7.57e-07
Sexo						
Hombre	.0281612	.0013773	20.45	0.000	.0254617	.0308606
Zona						
Bogotá	.0229699	.0017141	13.40	0.000	.0196105	.0263294

Note: dy/dx for factor levels is the discrete change from the base level.

Figura 7.5: Efecto marginal usando margins.

7.1.3 El modelo probit

El modelo probit, igual que el logit ya estudiado, aparece para solucionar los dos problemas fundamentales asociados al modelo de probabilidad lineal: el primero, que algunas probabilidades estimadas no se encuentren entre 0 y 1; y el segundo, que el efecto que tiene la variable explicatoria sobre la probabilidad de un evento particular sea constante, independientemente del valor que tome la variable explicatoria. El modelo probit, igual que el logit, garantiza que las probabilidades estén entre 0 y 1, pues utiliza la distribución de probabilidad acumulada de una variable aleatoria con función de densidad normal. Por otro lado, al utilizar el probit la función de distribución acumulada, la cual tiene forma de s, no permite que el efecto que tiene una variable explicatoria sobre la probabilidad sea constante, ya que en estos modelos hay tramos en los cuales la probabilidad de que se presente el evento de interés crece de forma muy lenta, principalmente cuando la variable explicatoria toma valores muy bajos o muy altos; mientras existen otros tramos donde la probabilidad de que se presente el evento crece muy rápidamente.

Para ilustrar la estimación del probit no agrupado utilizaremos los datos empleados en el modelo de probabilidad lineal y el logit en las subsecciones anteriores, es decir, se trata de identificar los determinantes de ser profesional, tomando como variable dependiente una variable dicótoma que toma el valor de 1 si la persona es profesional y 0 si no es profesional, y como variables explicatorias se tienen: el ingreso familiar; la zona, que toma el valor de 1 si reside en Bogotá y 0 si reside en otra zona del país; y el sexo, que toma el valor de 1 si es hombre y 0 si es mujer.

Para la estimación del probit se sigue esta ruta por ventanas:

Ventana

statistics → binary outcomes → probit regression → definir variable dependiente e Independientes → Ok.

Por comando sería:

Comando

La estimación del probit para el caso de estudio por comando será:

Comando

probit Profesional Ingresofamiliar Sexo Zona

. probit Profesional Ingresofamiliar Sexo Zona

```
Iteration 0: log likelihood = -5858.1761
Iteration 1: log likelihood = -5500.7817
Iteration 2: log likelihood = -5457.3903
Iteration 3: log likelihood = -5457.3173
Iteration 4: log likelihood = -5457.3173
```

```
Probit regression                                         Number of obs      =      55757
                                                               LR chi2(3)        =     801.72
                                                               Prob > chi2       =     0.0000
Log likelihood = -5457.3173                           Pseudo R2        =     0.0684
```

Profesional	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Ingresofamiliar	9.28e-06	5.87e-07	15.80	0.000	8.13e-06 .0000104
Sexo	.5778072	.0271983	21.24	0.000	.5244996 .6311148
Zona	.3876152	.0368379	10.52	0.000	.3154143 .4598161
cons	-2.75657	.0407913	-67.58	0.000	-2.83652 -2.676621

Figura 7.6: Estimación del modelo probit.

La Figura 7.6 muestra el resultado obtenido para el modelo probit, dichos resultados son similares a los del modelo logit, en términos de su significancia y el signo, implicando este último el efecto positivo sobre la probabilidad de ser profesional, del ingreso familiar, la zona y el sexo.

Después de estimar el modelo probit se pueden hallar los efectos marginales para un conjunto de valores representativos, es decir, los cambios en probabilidad ante cambios en las variables explicatorias.

. mfx, at(Ingresofamiliar =200 Sexo=1 Zona=1)

```
Marginal effects after probit
y = Pr(Profesional) (predict)
= .03678386
```

variable	dy/dx	Std. Err.	z	P> z	[95% C.I.]	x
Ingres-r	7.47e-07	.00000	15.87	0.000	6.5e-07 8.4e-07	200
Sexo*	.0278198	.00136	20.42	0.000	.02515 .03049	1
Zona*	.0220401	.00171	12.90	0.000	.018692 .025389	1

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Figura 7.7: Efectos marginales del probit.

La [Figura 7.7](#) muestra los efectos marginales que se obtienen con el Probit, indicando que cuando el ingreso familiar pasa de 200,000 pesos a 201,000 pesos, la probabilidad de ser profesional se incrementa en tan solo 0.0000747%, mientras ser hombre incrementa dicha probabilidad en 2.78% en relación a la mujer, y residir en Bogotá incrementa la probabilidad de ser profesional en 2.20% con relación a un no residente en Bogotá.

Es importante aclarar que el comando **margins** también permite calcular los efectos marginales en un probit, pero para hacerlo adecuadamente, en la estimación del modelo probit las variables binarias deben señalarse anteponiendo la letra **i.**, como se muestra a continuación:

Comando

```
probit Profesional Ingresofamiliar i.Sexo i.Zona
```

```
. probit Profesional Ingresofamiliar i.Sexo i.Zona
```

```
Iteration 0: log likelihood = -5858.1761
Iteration 1: log likelihood = -5500.7817
Iteration 2: log likelihood = -5457.3903
Iteration 3: log likelihood = -5457.3173
Iteration 4: log likelihood = -5457.3173
```

Probit regression	Number of obs	=	55757
	LR chi2(3)	=	801.72
	Prob > chi2	=	0.0000
Log likelihood = -5457.3173	Pseudo R2	=	0.0684

Profesional	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Ingresofamiliar	9.28e-06	5.87e-07	15.80	0.000	8.13e-06 .0000104
Sexo					
Hombre	.5778072	.0271983	21.24	0.000	.5244996 .6311148
Zona					
Bogotá	.3876152	.0368379	10.52	0.000	.3154143 .4598161
_cons	-2.75657	.0407913	-67.58	0.000	-2.83652 -2.676621

Figura 7.8: Estimacion probit para margins.

La [Figura 7.8](#) muestra que las estimaciones del probit anteponiendo la **i.** a las variables cualitativas coinciden con las obtenidas con el método estándar. Para calcular los efectos marginales del probit a través del **margins** se utiliza el siguiente comando:

Comando

```
margins, dydx(*) at(Ingresofamiliar=200 Sexo=1 Zona=1)
```

. margins, dydx(*) at(Ingresofamiliar =200 Sexo=1 Zona=1)								
Conditional marginal effects				Number of obs		= 55,757		
Model VCE : OIM								
 Expression : Pr(Profesional), predict()								
dy/dx w.r.t. : Ingresofamiliar 1.Sexo 1.Zona								
at : Ingresofamiliar = 200								
Sexo = 1								
Zona = 1								
 <hr/>								
Delta-method								
dy/dx Std. Err. z P> z [95% Conf. Interval]								
Ingresofamiliar	7.47e-07	4.70e-08	15.87	0.000	6.54e-07	8.39e-07		
Sexo								
Hombre	.0278198	.0013621	20.42	0.000	.0251501	.0304895		
Zona								
Bogotá	.0220401	.0017085	12.90	0.000	.0186916	.0253887		

Note: dy/dx for factor levels is the discrete change from the base level.

Figura 7.9: Efecto marginal con margins.

Como se muestra en la Figura 7.9, el cálculo de los efectos marginales en un probit para un conjunto de valores representativos usando margins, coinciden con los obtenidos a través del comando mfx.

Finalmente, una pregunta que surge frecuentemente en los cursos de econometría es: ¿Cuál de los dos modelos es mejor, logit o probit? La respuesta es que hasta el momento no existe un argumento de peso para decidir entre los dos modelos, más aún cuando ambos modelos arrojan resultados muy similares, aunque es claro que en ocasiones existen diferencias, principalmente cuando las probabilidades son calculadas en las colas. Una solución para seleccionar entre los dos modelos es utilizar los criterios de información de Akaike y Schwartz presentados en el capítulo 4.

7.1.4 Modelos de datos de recuento

Una variable dependiente discreta no negativa que toma pocos valores se denomina variable de recuento, como ejemplos de dichas variables se tienen: número de libros leídos en un año, número de visitas al dentista en un año, número de accidentes automovilísticos en un año, etc. Los modelos de regresión de Poisson, de regresión exponencial y de regresión binomial negativa son apropiados para modelar este tipo de fenómenos.

Debido a que la variable dependiente suele tomar el valor de 0 en estos modelos, no se puede aplicar una transformación logarítmica, entonces se debe modelar $E(Y | X)$ eligiendo algunas formas funcionales que aseguren que las predicciones para la variable dependiente Y sean también positivas.

7.1.5 Modelo de regresión de Poisson

El modelo de regresión de Poisson se ha utilizado para estudiar datos de recuento; en este modelo se asume que cada Y_i es una realización de una variable aleatoria con distribución Poisson, y que la esperanza condicional de Y dado X, se ajusta a una función exponencial, como se muestra en la ecuación (7.2):

$$E(Y | X_1, \dots, X_k) = \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k) = \exp(\mathbf{X}\boldsymbol{\beta}) \quad (7.2)$$

Tomando logaritmo a los dos lados se tiene la ecuación (7.3):

$$\ln(E(Y | X_1, \dots, X_k)) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k = \mathbf{X}\boldsymbol{\beta} \quad (7.3)$$

Por lo tanto, el logaritmo del valor esperado condicional de Y dado X es una función lineal. Asumiendo que la media de la distribución de

Poisson, que es la que la caracteriza, es igual a $\exp(\mathbf{X}\beta)$, se puede calcular la probabilidad de que Y tome el valor de h , condicionada a valores dados a las variables explicatorias \mathbf{X} , como se muestra en la [ecuación \(7.4\)](#):

$$P(Y = h | \mathbf{X}) = \frac{\exp(-\exp(\mathbf{X}\beta))(\exp(\mathbf{X}\beta))^h}{h!} \quad (7.4)$$

Como se trata en los cursos de Estadística básica, las probabilidades y los momentos de orden superior en una distribución de Poisson están determinados por la media; es decir, la varianza condicional de Y dado \mathbf{X} , $V(Y | \mathbf{X})$ es igual a la media condicional de Y dado \mathbf{X} , $E(Y | \mathbf{X})$, lo cual en ocasiones no se cumple en los modelos de Poisson. No obstante la limitación anterior, la distribución de Poisson es robusta en el sentido de que, aunque la distribución no sea estrictamente Poisson, es factible encontrar estimadores de los coeficientes de regresión del modelo Poisson que sean consistentes y asintóticamente normales (Wooldridge, 2010).

Cuando Y dado X_1, \dots, X_k no tiene una distribución exactamente de Poisson, la estimación máxima verosímil se denomina estimación por quasi máxima verosimilitud (**CMV**). Cuando estimamos por **CMV** y no se cumple el supuesto de $E(Y | \mathbf{X}) = V(Y | \mathbf{X})$ para poder calcular los errores estándar, se debe suponer que la varianza condicional es proporcional a la media condicional, como se muestra en la [ecuación \(7.5\)](#):

$$V(Y | \mathbf{X}) = \sigma^2 E(Y | \mathbf{X}) \quad (7.5)$$

Donde $\sigma^2 > 0$ es un parámetro desconocido. Si $\sigma^2=1$, tenemos el supuesto sobre la varianza de la distribución Poisson. Si $\sigma^2 > 1$, tenemos sobre-dispersión, que es lo que sucede en la mayoría de aplicaciones. Si $\sigma^2 < 1$, tenemos infradispersión, lo cual se presenta con menor frecuencia en el trabajo aplicado (Wooldridge, 2010).

Para ilustrar el modelo de regresión de Poisson se utilizan datos tomados de la encuesta de consumo cultural realizada por el DANE, donde se tomará como variable dependiente el número de libros leídos por una persona en un trimestre (**numlib_3m2**) y como variables independientes:

- La actividad principal, con seis categorías (1. trabajar, 2. estudiar, 3. oficios del hogar, 4. buscar trabajo, 5. incapacitado y 6. otras actividades).
- El sexo, con dos categorías (0. femenino y 1. masculino).
- El estrato, con tres categorías (1. bajo, 2. medio y 3. alto).
- La participación en actividades culturales, con dos categorías (0. no participa y 1. sí participa).
- Lugar de residencia, con tres categorías (1. si reside en resto de cabecera, 2. si reside en una de las diez ciudades intermedias y 3. si reside en una de las trece principales ciudades).
- Nivel educativo, con cuatro categorías (2. si tiene educación primaria, 3. si tiene educación secundaria, 4. si tiene educación superior y 5. si tiene educación a nivel de posgrado).

Para la estimación por ventanas del modelo de regresión de Poisson se deben seguir estos pasos:

Ventana

1. statistics → count outcomes → Poisson regression.
2. Colocar la variable dependiente que es número de libros leídos en un mes.
3. Colocar las variables Independientes como las variables escogidas por el investigador (por ejemplo: las variables mencionadas anteriormente).

NOTA: Dentro de las opciones de la estimación se debe dirigir a la ventana con nombre SE → Robust y colocar la opción **Robust**, esto se hace debido a que el modelo de Poisson es heterocedástico y con los estimadores robustos de White se ajustan los errores estándar de los estimadores.

La estimación del modelo Poisson por comando es:

Comando

```
poisson variable_dependiente_variables_independientes, robust
```

La [Figura 7.10](#) muestra la estimación Poisson robusta para el número de libros leídos.

. poisson numlib_3m2 i.Act_Ppal i.Sexo i.Parti_Cult i.Lugar i. Niv_Educ, robust						
Iteration 0: log pseudolikelihood = -3778.1269						
Iteration 1: log pseudolikelihood = -3778.1268						
Poisson regression						
Number of obs = 2398						
Wald chi2(12) = 138.96						
Prob > chi2 = 0.0000						
Log pseudolikelihood = -3778.1268						
Pseudo R2 = 0.0239						
		Robust				
numlib_3m2		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Act_Ppal						
Estudiar		-.057645	.0961228	-0.60	0.549	-.2460421 .1307521
Oficios del hogar		.1604549	.0812401	1.98	0.048	.0012272 .3196825
Buscar trabajo		-.0044097	.0442964	-0.10	0.921	-.091229 .0824096
Incapacitado		-.0008633	.1210547	-0.01	0.994	-.238126 .2363995
Otra actividad		.1414529	.081782	1.73	0.084	-.0188368 .3017426
Sexo						
Masculino		.119292	.0499382	2.39	0.017	.0214149 .2171692
Parti_Cult						
Si		.1989994	.0776507	2.56	0.010	.0468068 .351192
Lugar						
10 Ciudades Intermedias		-.0318306	.0582408	-0.55	0.585	-.1459805 .0823194
13 Principales Ciudades		.1011915	.0499399	2.03	0.043	.0033111 .199072
Niv_Educ						
Secundaria		.1514286	.0444623	3.41	0.001	.064284 .2385732
Superior		.3695658	.0565253	6.54	0.000	.2587784 .4803533
Posgrado		.5254561	.0825722	6.36	0.000	.3636175 .6872947
_cons		.1167391	.0519725	2.25	0.025	.0148749 .2186032

Figura 7.10: Estimación Poisson robusta.

La estimación anterior corresponde al valor medio estimado para el i -ésimo individuo. A partir de esta estimación es posible hallar el valor medio para un individuo con determinadas características, y con este valor medio se puede hallar la probabilidad de que dicho individuo lea cierto número de libros en el trimestre. Por otro lado, se observa que los coeficientes de regresión son significativos individualmente a un nivel del 10%. Aquellos que impactan negativamente la lectura de libros en el último trimestre son los asociados a las categorías: ser estudiante, estar buscando empleo y residir en alguna de las diez ciudades intermedias del país; mientras que los que impactan positivamente la lectura de libros en el último trimestre son los coeficientes asociados a las categorías: tener como actividad principal oficios del hogar u otra actividad, ser hombre, pertenecer al estrato socioeconómico medio, participar en actividades culturales, residir en alguna de las trece principales ciudades del país y tener un nivel educativo de secundaria, superior o posgrado. La interpretación de los coeficientes de regresión es similar a la de los modelos lineales en su forma funcional log-lin, presentados en el [capítulo 2](#) de este manual.

A través del cálculo se demuestra que los efectos marginales, que miden el cambio en el valor medio de Y ante un cambio unitario en la j -ésima variable explicatoria, vienen dados por el producto entre el j -ésimo coeficiente y el valor esperado estimado de Y condicionado a un conjunto de valores asignados a las variables explicatorias, por lo cual este valor esperado dependerá de la evaluación en la función de regresión estimada de los valores asignados a las variables explicatorias. Los efectos marginales en el modelo de Poisson pueden calcularse a través del comando **margins**, ya comentado en los modelos logit y probit:

Comando

`margins, dydx(*)`

Modelo de regresión binomial negativa

Este modelo se utiliza para casos de sobredispersión, ya que supone que:

$$V(Y | \mathbf{X}) = \sigma^2$$

$$E(Y | \mathbf{X}) = (1 + \eta^2)E(Y | X)$$

La estimación de los parámetros β , η^2 se realiza por el método de máxima verosimilitud. Para que las estimaciones sean consistentes y eficientes se tiene que cumplir el supuesto de binomial negativa.

Una prueba para determinar si existe sobredispersión, donde la hipótesis nula es que no existe sobredispersión versus la hipótesis alterna que existe sobredispersión, está basada en la [ecuación \(7.6\)](#):

$$V(Y | \mathbf{X}) = \sigma^2 E(Y | \mathbf{X}) = (1 + \eta^2)E(Y | X) \quad (7.6)$$

La hipótesis nula $H_0 : \alpha = 0$ versus $H_1 : \alpha > 0$.

Para la estimación por ventanas del modelo de regresión de binomial negativa se deben seguir estos pasos:

Ventana

1. statistics → count outcomes → Negative binomial regression.
2. Colocar la variable dependiente que es número de libros leídos en un mes y en las Independientes las variables escogidas por cada Investigador.

NOTA: En las opciones de la estimación se debe dirigir a la ventana con nombre SE → Robust y se coloca la opción **Robust**, esto se hace debido a que este modelo presenta heterocedasticidad y con los estimadores robustos de White se ajusta el error estándar de los estimadores.

Por comando, La estimación binomial negativa es:

Comando

```
nbreg variable_dependiente_variables_independientes, robust
```

La [Figura 7.11](#) muestra la estimación binomial negativa robusta para el número de libros leídos.

Negative binomial regression		Number of obs = 2398				
Dispersion = mean		Wald chi2(12) = 137.88				
Log pseudolikelihood = -3691.0038		Prob > chi2 = 0.0000				
<hr/>						
numlib_3m2		Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
Act_Fpal						
Estudiar	-0.0567938	.0951802	-0.60	0.551	-.2433437	.129756
Oficios del hogar	.1637903	.0808425	2.03	0.043	.0053419	.3222388
Buscar trabajo	-.0043029	.0438039	-0.10	0.922	-.0901569	.0815511
Incapacitado	.0020572	.1197608	0.02	0.986	-.2326697	.2367841
Otra actividad	.140304	.0812302	1.73	0.084	-.0189043	.2995123
Sexo						
Masculino	.1189631	.0490758	2.42	0.015	.0227764	.2151499
Parti_Cult						
Si	.1990858	.0775115	2.57	0.010	.0471662	.3510055
Lugar						
10 Ciudades Intern	-.0315223	.0570223	-0.55	0.580	-.1432841	.0802394
13 Principales Ciudades	.1009096	.0490038	2.06	0.039	.004864	.1969553
Niv_Educ						
Secundaria	.1493608	.0439229	3.40	0.001	.0632735	.2354482
Superior	.3698615	.0560824	6.59	0.000	.259942	.479781
Posgrado	.5254181	.0823264	6.38	0.000	.3640614	.6867748
_cons	.1172295	.0516466	2.27	0.023	.016004	.2184551
/lnalpha	-2.087811	.346375			-2.766693	-1.408928
alpha	.1239582	.042936			.0628695	.2444051

Figura 7.11: Estimación binomial negativa robusta.

Hay métodos desarrollados para concluir si se debe usar regresión de Poisson o binomial negativa con una base de datos, en este manual utilizaremos el proporcionado por **Stata** al estimar el modelo binomial negativo. Este test viene incorporado en la parte de abajo de la estimación de binomial negativa, en esta aparece la variable **alpha** con su valor estimado, en nuestro caso es 0.12, la significancia de dicho valor es contrastado a través de la hipótesis nula $H_0 : \alpha = 0$. Para probar esta hipótesis se utiliza el test de razón de verosimilitud. El valor **P** arrojado en esta prueba es igual a 0.0000, lo que indica que se rechaza la hipótesis nula y, por tanto, existe sobredispersión y la binomial negativa es adecuada. Finalmente, los coeficientes de regresión y el efecto marginal en la binomial negativa se interpretan de la misma manera que en el modelo Poisson, porque ambos modelos tienen la misma media condicional.

7.1.6 El modelo de regresión Tobit

El modelo de Tobit puede entenderse como una extensión del modelo probit, pues en este último se proporciona la probabilidad de que se presente un evento determinado, mientras en el modelo de Tobit se está interesado en modelar una variable que se desprende del modelo probit, la cual presenta muchos ceros, y los valores que no son cero son muy variados y siguen una distribución aproximadamente continua. Para entender lo anterior, un ejemplo es cuando estamos interesados en encontrar los determinantes de la tenencia de vehículo propio, en este caso utilizamos un modelo probit; posteriormente nos interesa conocer los determinantes de los gastos del vehículo, cuando se selecciona la muestra es posible que no todos los entrevistados tengan un vehículo, pero sí se tiene la información sobre las variables independientes, por tanto, lo ideal es no dejar a un lado las personas que no tengan vehículos y que por consiguiente su gasto es 0. En esta situación no se debe estimar el modelo por **MCO**, pues es posible obtener valores estimados negativos, lo que producirá predicciones negativas. Adicionalmente, las estimaciones **MCO** son sesgadas e inconsistentes, como lo muestra (Wooldridge, 2010). El modelo Tobit resulta ser el apropiado para este tipo de situaciones y debe ser estimado por máxima verosimilitud o por el método propuesto por (Heckman, 1977), el cual consiste en dos pasos: primero se estima la probabilidad de que el individuo tenga vehículo usando un modelo probit, posteriormente se estima el modelo de gasto en vehículo adicionándole un término denominado

razón de Mills, el cual se desprende de la estimación del modelo probit. Actualmente, para la estimación del Tobit se emplea el método de máxima verosimilitud, dados los desarrollos computacionales y la mayor eficiencia de este método frente a la utilización de razón de Mills.

Para ilustrar el modelo Tobit se utilizarán los datos del archivo **MROZ** de la base de datos de (Wooldridge, 2010), dicho archivo contiene información sobre: *hours*, horas trabajadas en 1975; *kidslt6*, niños menores de 6 años; *kidsge6*, niños entre 6 y 18 años; *age*, edad en años; *educ*, años de escolaridad; *exper*, experiencia laboral en años; *expert*, experiencia laboral al cuadrado en años; y *nwifeinc*, índice de ingreso no laboral= (ingreso familiar-ingreso laboral)/1000, en 1975. La muestra está conformada por 753 mujeres casadas, de las cuales 428 trabajaron fuera de casa para obtener un salario y 325 no trabajaron ni una hora. El número de horas trabajadas oscila entre 12 y 4950. El problema que se plantea es encontrar los determinantes del número de horas trabajadas fuera de casa de las mujeres casadas, como 325 mujeres no trabajaron ni una hora, en la variable dependiente reciben el valor de 0, por lo tanto, la estructura de la información se adapta a la aplicación del modelo Tobit.

Las [Figuras 7.12](#) y [7.13](#) muestran la estimación del modelo utilizando el método Tobit y los MCO. Como se presenta en los cursos de Econometría básica, los coeficientes estimados por estos métodos no son directamente comparables, para compararlos como lo sugiere (Wooldridge, 2010) se deben ajustar los coeficientes estimados por el método Tobit, como se muestra en la [ecuación \(7.7\)](#):

$$\frac{\partial E(Y | X)}{\partial X_j} = \beta_j \Phi\left(\frac{X\beta}{\sigma}\right) \quad (7.7)$$

```
. tobit hours nwifeinc educ exper expersq age kidsge6 kidslt6, ll(0)
```

Tobit regression

Number of obs = 753

LR chi2(7) = 271.59

Prob > chi2 = 0.0000

Log likelihood = -3819.0946

Pseudo R2 = 0.0343

hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
nwifeinc	-8.814243	4.459096	-1.98	0.048	-17.56811 -.0603724
educ	80.64561	21.58322	3.74	0.000	38.27453 123.0167
exper	131.5643	17.27938	7.61	0.000	97.64231 165.4863
expersq	-1.864158	.5376615	-3.47	0.001	-2.919667 -.8086479
age	-54.40501	7.418496	-7.33	0.000	-68.96862 -39.8414
kidsge6	-16.218	38.64136	-0.42	0.675	-92.07675 59.64075
kidslt6	-894.0217	111.8779	-7.99	0.000	-1113.655 -674.3887
_cons	965.3053	446.4358	2.16	0.031	88.88528 1841.725
/sigma	1122.022	41.57903			1040.396 1203.647

```
Obs. summary: 325 left-censored observations at hours<=0
               428 uncensored observations
               0 right-censored observations
```

Figura 7.12: Estimación Tobit.

```
. regress hours nwifeinc educ exper expersq age kidsge6 kidslt6
```

Source	SS	df	MS	Number of obs = 753		
Model	151647606	7	21663943.7	F(7, 745) = 38.50		
Residual	419262118	745	562767.944	Prob > F = 0.0000		
Total	570909724	752	759188.463	R-squared = 0.2656		
				Adj R-squared = 0.2587		
				Root MSE = 750.18		
hours	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nwifeinc	-3.446636	2.544	-1.35	0.176	-8.440898	1.547626
educ	28.76112	12.95459	2.22	0.027	3.329283	54.19297
exper	65.67251	9.962983	6.59	0.000	46.11365	85.23138
expersq	-.7004939	.3245501	-2.16	0.031	-1.337635	-.0633524
age	-30.51163	4.363868	-6.99	0.000	-39.07858	-21.94469
kidsge6	-32.77923	23.17622	-1.41	0.158	-78.2777	12.71924
kidslt6	-442.0899	58.8466	-7.51	0.000	-557.6148	-326.565
_cons	1330.482	270.7846	4.91	0.000	798.8906	1862.074

Figura 7.13: Estimación Tobit por MCO.

Donde el factor de ajuste $\Phi(\frac{\hat{\alpha} - \beta}{\sigma})$ se consigue evaluando en la función de distribución acumulada normal estándar el valor ajustado en los valores medios de las variables explicatorias, dividido entre la desviación estándar estimada, en nuestro caso, $\hat{\sigma} = 1122.02$. Así, por ejemplo, un año adicional de educación incrementa el número de horas trabajadas en el año, en promedio en $0,6045 * 80,65$, *ceteris paribus*, lo cual es bastante distinto al valor estimado por el método MCO. De la misma manera se realiza para los demás coeficientes de regresión del modelo estimado por el método Tobit. Esto implica que el cálculo de la desviación estándar con el método Tobit es supremamente importante, pues es a partir de este que se realizan los ajustes.

En ocasiones no estamos seguros si el modelo Tobit es adecuado; una manera de validarla es estimar un modelo probit en el cual la variable dependiente toma el valor de 1 si la variable dependiente Y toma un valor mayor que 0, y de 0 si la variable dependiente Y toma un valor igual a 0, si el j -ésimo coeficiente estimado y_j del modelo probit está cercano al j -ésimo coeficiente estimado del modelo Tobit dividido entre la desviación estándar del error, entonces el modelo Tobit es adecuado. La Figura 7.14 muestra las estimaciones del modelo probit, por lo tanto, al comparar el coeficiente de experiencia estimado con el modelo probit 0.1233, con el cociente del coeficiente estimado con el modelo Tobit para la variable experiencia, dividido entre $\hat{\sigma}$, $(131,56/1122,02) = 0,1172$, se observa que los dos valores son cercanos. Repitiendo este proceso para todos los demás coeficientes, se puede verificar que el modelo Tobit para el caso estudiado es adecuado.

. probit inlf nwifeinc educ exper expersq age kidsge6 kidslt6						
Iteration 0: log likelihood = -514.8732						
Iteration 1: log likelihood = -402.06651						
Iteration 2: log likelihood = -401.30273						
Iteration 3: log likelihood = -401.30219						
Iteration 4: log likelihood = -401.30219						
Probit regression				Number of obs = 753		
				LR chi2(7) = 227.14		
				Prob > chi2 = 0.0000		
Log likelihood = -401.30219				Pseudo R2 = 0.2206		
inlf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
nwifeinc	-.0120237	.0048398	-2.48	0.013	-.0215096	-.0025378
educ	.1309047	.0252542	5.18	0.000	.0814074	.180402
exper	.1233476	.0187164	6.59	0.000	.0866641	.1600311
expersq	-.0018871	.0006	-3.15	0.002	-.003063	-.0007111
age	-.0528527	.0084772	-6.23	0.000	-.0694678	-.0362376
kidsge6	.036005	.0434768	0.83	0.408	-.049208	.1212179
kidslt6	-.8683285	.1185223	-7.33	0.000	-1.100628	-.636029
_cons	.2700768	.508593	0.53	0.595	-.7267473	1.266901

Figura 7.14: Estimación probit.

Modelos de regresión con datos de panel

Temas Tratados

- [8.1 Modelo de regresión con MCO agrupados o de coeficientes constantes](#)
- [8.2 Modelo de efectos fijos usando variable dummy](#)
- [8.3 Modelo de efectos aleatorios \(MEFA\)](#)
- [8.4 Estimación de un modelo con datos de panel en Stata](#)
 - [8.4.1 Modelo de panel de coeficientes constantes](#)
 - [8.4.2 Modelo de efectos aleatorios \(MEFA\)](#)
 - [8.4.3 Modelo de panel de mínimos cuadrados con variable dicotómica de efectos fijos](#)
 - [8.4.4 Test de Hausman para escoger entre efectos fijos y aleatorios](#)
- [8.5 Apéndice](#)

Los modelos de regresión con datos de panel son complejos, es por ello que en el presente capítulo se abordan su importancia y su aplicación de una manera introductoria, siguiendo (Gujarati y Porter, 2010), con el fin de que el lector tenga una visión general de estos modelos.

Ahora bien, hasta el momento los modelos estudiados han hecho uso de datos de corte transversal y de series de tiempo, sin embargo, existen situaciones en las cuales es posible diseñar modelos donde se utilice la combinación de ambos tipos de datos, que se obtienen a partir del seguimiento a las mismas unidades de corte transversal a lo largo del tiempo. Varias son las ventajas de los datos de panel, como lo plantean (Gujarati y Porter, 2010); una de ellas es que permiten estudiar la heterogeneidad inobservable que se presenta entre las unidades (empresas, individuos, países, ciudades, regiones, etc.). Al combinar las series de tiempo con las observaciones de corte transversal, el número de observaciones se incrementa significativamente, disminuyendo así la posible presencia de multicolinealidad. Es posible estudiar la dinámica que presentan a lo largo del tiempo dichas unidades. En resumen, los datos de panel enriquecen el análisis empírico, que no sería posible con datos solo de corte transversal o solo de series de tiempo. Los datos de panel pueden ser balanceados, es decir, cuando se tiene información completa para todas las unidades, en caso contrario, el dato de panel es no balanceado.

Aunque en los textos, por lo general, se presentan cuatro posibilidades de estimar un panel, en este manual se consideran solo tres, a saber; modelo Pool o de coeficientes constantes, modelo de efectos fijos y modelo de efectos aleatorios.

8.1 Modelo de regresión con MCO agrupados o de coeficientes constantes

Considere un modelo con tres variables explicatorias como se muestra en la [ecuación \(8.1\)](#):

$$Y_{it} = \beta_1 + \beta_2 X_{1it} + \beta_3 X_{2it} + \beta_4 X_{3it} + U_{it} \quad (8.1)$$

Donde se asume: $i = 1, 2, \dots, 6$ y $t = 1, 2, \dots, 15$. i representa la i -ésima observación, unidad o empresa considerada y t es el periodo a lo largo del cual se hace el seguimiento a dichas unidades. Y es la variable dependiente y X_1, X_2 y X_3 son las variables explicatorias, las cuales se asumen no aleatorias en muestreos repetidos y, de serlo, son exógenas estrictamente, es decir, no están correlacionadas con los errores contemporáneos, pasados y futuros. Finalmente se asume que los errores son independientes e idénticamente distribuidos con media 0 y varianza constante.

La estimación por MCO agrupada asume que los impactos que tienen las variables explicatorias medidos a través de las pendientes, así como el intercepto, son comunes a todas las unidades, es decir, se considera que no existe heterogeneidad entre las unidades; de existir heterogeneidad medida por una variable no observable, esta se estaría omitiendo e iría a parar al error, lo cual generaría que los estimadores MCO fuesen sesgados e inconsistentes, lo que significa que los valores esperados de los estimadores no serán iguales a los parámetros poblacionales, y, aunque el tamaño de muestra crezca, los estimadores no van a tender a los parámetros poblacionales.

8.2 Modelo de efectos fijos usando variable dummy

El supuesto de que no existe distinción entre las observaciones, unidades o empresas, es difícil de sostener, por tanto, se debe plantear un modelo que considere la presencia de heterogeneidad entre las unidades. El modelo de efectos fijos permite que cada empresa tenga su propio intercepto, y esto se logra incorporando variables dicotómicas cuyos coeficientes miden los diferenciales en intercepto entre las empresas. El nombre

de efecto fijo se debe a que si tal diferencial en el intercepto existe, este se mantiene fijo a lo largo del tiempo. A continuación se considera el modelo anterior, pero incluyendo diferentes interceptos, como se ve en la [ecuación \(8.2\)](#):

$$Y_{it} = \beta_{1i} + \beta_2 X_{1it} + \beta_3 X_{2it} + \beta_4 X_{3it} + U_{it} \quad (8.2)$$

El subíndice i en el intercepto del modelo anterior significa que los interceptos de las unidades o empresas pueden ser diferentes. Las diferencias quizás se deban a características propias de cada unidad o empresa, como el estilo de dirección, la filosofía, la administración, etc.

En la [ecuación \(8.3\)](#) se muestra una forma alterna de presentar el modelo anterior. El nuevo modelo incorpora $m-1$ variables **dummy**, donde m representa las unidades o empresas transversales. Se incorporan $m-1$ variables **dummy** para evitar multicolinealidad perfecta; en nuestro caso, como se tienen seis empresas, se incorporan solo cinco. Los coeficientes de estas **dummy**, como ya se dijo, miden el diferencial en los interceptos con relación a la categoría base o empresa 1:

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \alpha_6 D_{6i} + \beta_2 X_{1it} + \beta_3 X_{2it} + \beta_4 X_{3it} + U_{it} \quad (8.3)$$

Donde D_2 es igual a 1, si la observación corresponde a la empresa o unidad 2, y 0 en otro caso; D_3 es igual a 1 si la observación corresponde a la empresa 3, y 0 en otro caso; y así sucesivamente.

Para determinar si existe o no heterogeneidad entre las unidades o empresas se plantean las siguientes hipótesis.

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$$
$$H_1 : \text{Al menos un } \alpha_i \neq 0$$

El estadístico utilizado para probar la hipótesis anterior es un **F** de mínimos cuadrados restringidos, como se ve en la [ecuación \(8.4\)](#), el cual ya se ha presentado en secciones anteriores:

$$F = \frac{(SCR_0 - SCR_a)/m}{SRC_a/(n - k)}, \quad (8.4)$$

Donde:

- SCR_0 = suma de cuadrados residuales bajo la hipótesis nula o restringida.
- SCR_a = suma de cuadrados residuales bajo la hipótesis alterna o no restringida.
- m = número de restricciones o el número de coeficientes de regresión que desaparecen del modelo original al imponer la restricción.

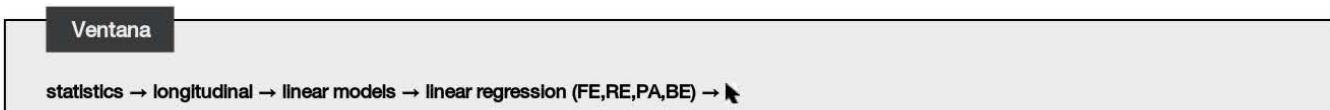
Si se rechaza la hipótesis nula, entonces el modelo adecuado es el de efectos fijos. Una forma alterna de probar la hipótesis anterior se encuentra incorporada en el Stata a través de los siguientes pasos:

1. Declare inicialmente sus datos como datos de panel a través de la siguiente ruta:



Defina la variable que identifica su unidad de observación, empresas o individuos, y su variable que identifica la variable tiempo, así como la frecuencia de los datos.

2. Despues de declarar sus datos como datos de panel, estime el modelo de efectos fijos siguiendo la ruta:



Defina la variable dependiente y sus variables independientes o explicatorias y active la opción Fixed effects → enter. El paquete arrojará en la parte inferior un valor **F**, asociado al test donde todos los coeficientes de regresión asociados a las variables **dummy** sean 0.

Usando la línea de comando también se puede estimar un modelo de efectos fijos:



Es importante aclarar que el modelo de efectos fijos considerado es unidireccional, porque se está asumiendo que existe un solo efecto diferencial y que este es atribuible a las unidades o empresas; no obstante, podría también existir un efecto atribuible al tiempo y, en este caso, el modelo de efectos fijos sería bidireccional. El modelo de efectos fijos bidireccional o unidireccional usando variables dummy o dicótomas presenta algunas desventajas, como:

- Exceso de variables dicótomas si el número de individuos N fuese grande, por consiguiente, no habrá observaciones suficientes para un análisis significativo.
- Posible presencia de multicolinealidad, dada la cantidad de variables dummy, lo que puede dificultar la estimación precisa de algunos coeficientes de regresión.
- Es posible que no identifiquen el efecto de variables que no cambian con el tiempo.

8.3 Modelo de efectos aleatorios (MEFA)

Considere el modelo mostrado en la [ecuación \(8.5\)](#):

$$Y_{it} = \beta_1 i + \beta_2 X_{1it} + \beta_3 X_{2it} + \beta_4 X_{3it} + U_{it} \quad (8.5)$$

En este modelo se asume que las unidades seleccionadas son una muestra extraída de una población con una media común a todas las unidades que la componen, entonces los diferenciales que se presentan entre los interceptos son variables aleatorias, pues los mismos cambiarán cuando las unidades seleccionadas aleatoriamente de la población sean distintas. De acuerdo con lo anterior, $\beta_1 i = \beta_1 + e_i$, siendo β_1 la media común y e_i , el i -ésimo diferencial en el intercepto, que es una variable aleatoria. Reformulando el modelo anterior, se obtienen las [ecuaciones \(8.6\)](#), [\(8.7\)](#) y [\(8.8\)](#):

$$Y_{it} = \beta_1 + e_i + \beta_2 X_{1it} + \beta_3 X_{2it} + \beta_4 X_{3it} + U_{it} \quad (8.6)$$

$$Y_{it} = \beta_1 + \beta_2 X_{1it} + \beta_3 X_{2it} + \beta_4 X_{3it} + e_i + U_{it} \quad (8.7)$$

$$Y_{it} = \beta_1 + \beta_2 X_{1it} + \beta_3 X_{2it} + \beta_4 X_{3it} + W_{it} \quad (8.8)$$

De allí que este modelo también se conozca como modelo de componentes del error, porque el término de error W_{it} lo conforman dos términos, e_i , el error específico asociado al i -ésimo individuo, y u_{it} , el error de la i -ésima unidad o individuo en el tiempo t .

La racionalidad detrás del modelo de efectos aleatorios, a diferencia del modelo de efectos fijos, se presenta cuando el efecto individual se asume que no está correlacionado con las variables explicatorias en el modelo. Una ventaja del modelo de efectos aleatorios es que se pueden incluir variables que no cambian en el tiempo, como el género, dado que en el modelo de efectos fijos estas variables son absorbidas por el intercepto. Finalmente, la interpretación de los coeficientes en el modelo de efectos aleatorios es complicada, ya que incluyen tanto los efectos dentro de las unidades como entre las unidades.

La estimación de un modelo de panel de datos de efectos aleatorios se hace siguiendo la misma ruta utilizada para estimar un modelo de efectos fijos, así:

- Declare inicialmente sus datos como datos de panel a través de la siguiente ruta:



Defina la variable que identifica su unidad de observación, empresas, individuos, y su variable que identifica la variable tiempo, así como la frecuencia de los datos.

- Despues de declarar sus datos como datos de panel, estime el modelo de efectos aleatorios siguiendo la ruta:



Defina la variable dependiente y sus variables independientes o explicatorias y **Ok**, pues el programa trae por defecto **default** la estimación para efectos aleatorios.

La forma de estimar un modelo de efectos aleatorios a través de comandos es como sigue:

Comando

```
xtset Id year, yearly xtreg Y X1 X2 X3, re
```

8.4**Estimación de un modelo con datos de panel en Stata**

Caso de estudio. En una base pública de archivos de datos, se encuentra información quinquenal sobre el índice de Gini, escolaridad media (**yschooling**) y población urbana (**poburb**) para 39 países, entre los años 1980 y 1995. El problema propuesto es encontrar los determinantes de la desigualdad medida por el índice de Gini. Con base en las variables anteriores se formula el siguiente modelo (8.9).

$$\text{Gini}_{it} = \beta_1 + \beta_2(\text{yschoolin}_{it}) + \beta_3(\text{poburb}_{it}) + U_{it} \quad (8.9)$$

El Stata permite estimar de una manera sencilla los modelos de datos de panel, si la estructura de la base de datos está organizada en forma de *pooled* o apilonada.

8.4.1 Modelo de panel de coeficientes constantes

A continuación se presenta la estimación del modelo de datos de panel planteado, asumiendo coeficientes constantes. Recuerde, como se expuso en la [sección 8.1](#), que en este modelo se asume que no existe heterogeneidad entre las unidades y que, de existir, esta irá a parar al error, generando correlación entre el término de perturbación y algunas variables explicatorias, lo que hace que estimadores MCO sean sesgados e inconsistentes. La [Figura 8.1](#) muestra los resultados de la estimación del modelo de panel de coeficientes constantes a través del siguiente comando:

Comando

```
regress gini yschooling poburb
```

```
. regress gini yschooling poburb
```

Source	SS	df	MS	Number of obs	=	234
Model	5584.65405	2	2792.32703	F(2, 231)	=	43.55
Residual	14812.4861	231	64.1233167	Prob > F	=	0.0000
Total	20397.1402	233	87.5413742	R-squared	=	0.2738
				Adj R-squared	=	0.2675
				Root MSE	=	8.0077
gini	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yschooling	-2.430803	.2635204	-9.22	0.000	-2.950014	-1.911592
poburb	.1408597	.0316759	4.45	0.000	.0784492	.2032703
_cons	44.36969	1.805774	24.57	0.000	40.8118	47.92758

Figura 8.1: Modelo de panel con coeficientes constantes.

Al observar los resultados de la regresión agrupada y aplicar los criterios convencionales, se ve que todos los coeficientes de regresión son muy significativos estadísticamente. Como el término de error en este modelo incorpora la posible heterogeneidad existente, más el error idiosincrásico del modelo, se debe validar la presencia de autocorrelación en dicho error y su fuente a través del comando:

Comando

```
xtserial gini yschooling poburb, output
```

xtserial implementa una prueba de correlación serial en los errores idiosincrásicos de un modelo lineal de datos de panel. Esta prueba posee

buenas propiedades y una alta potencia para tamaños de muestra razonablemente grandes. La hipótesis en esta prueba es que no existe autocorrelación de orden 1.

Los resultados mostrados en la [Figura 8.2](#) indican presencia de autocorrelación en el error idiosincrático, pues el valor p es menor que un nivel de significancia del 5%.

. xtserial gini yschooling poburb, output						
Linear regression			Number of obs = 195			
D.gini	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
yschooling						
D1.	-.9460962	.5864807	-1.61	0.115	-2.133364	.2411718
poburb						
D1.	.1749776	.1102464	1.59	0.121	-.0482045	.3981597

(Std. Err. adjusted for 39 clusters in pais)

Wooldridge test for autocorrelation in panel data						
H0: no first order autocorrelation						
F(1, 38) =	13.596					
Prob > F =	0.0007					

Figura 8.2: Test de correlación serial.

8.4.2 Modelo de efectos aleatorios (MEFA)

Como se expresó en la primera parte de esta sección, los modelos de datos de panel de efectos aleatorios conciben el efecto diferencial en intercepto entre las diferentes unidades como aleatorio, eso obliga a llevar a estos diferenciales como un componente adicional del error del modelo, quedando dicho error compuesto por el error específico asociado al i -ésimo individuo, y_{it} , el error de la i -ésima unidad o individuo en el tiempo t . La [Figura 8.3](#), la cual muestra la estimación para el modelo de efectos aleatorios, se obtiene siguiendo la ruta por ventanas descrita arriba en la [sección 8.3](#) de modelos de efectos aleatorios o a través del comando:

Comando

```
regress gini yschooling poburb, re
```

Al comparar el modelo de panel agrupado con coeficientes constantes y el de efectos aleatorios, se observa que su estructura es similar, y si la varianza del error específico es igual a 0, no existiría diferencia entre ambos modelos y, por tanto, nos quedaríamos con el modelo *pool*. Para tomar la decisión sobre qué modelo utilizar, se realiza la prueba de multiplicador de Lagrange de Breusch-Pagan para efectos aleatorios. Si la hipótesis no se rechaza, el modelo de coeficientes constantes es el adecuado; en caso contrario, se elige el modelo de efectos aleatorios, el cual será contrastado con el de efectos fijos para seleccionar el adecuado.

. xtreg gini yschooling poburb, re	
Random-effects GLS regression	Number of obs = 234
Group variable: pais	Number of groups = 39
R-sq: within = 0.0085	Obs per group: min = 6
between = 0.2550	avg = 6.0
overall = 0.2083	max = 6
	Wald chi2(2) = 10.53
corr(u_i, X) = 0 (assumed)	Prob > chi2 = 0.0052
<hr/>	
gini	Coef. Std. Err. z P> z [95% Conf. Interval]
yschooling	-1.135576 .3499032 -3.25 0.001 -1.821373 -.4497779
poburb	.1232798 .0563576 2.19 0.029 .0128209 .2337387
_cons	36.41175 2.925142 12.45 0.000 30.67858 42.14492
sigma_u	7.0034547
sigma_e	3.9608998
rho	.75765479 (fraction of variance due to u_i)

Figura 8.3: Modelo de panel de datos de efectos aleatorios.

La prueba de Breusch-Pagan se realiza a través del comando:

Comando

xttest0

La Figura 8.4 muestra los resultados de la prueba de Breusch-Pagan, por tanto, se rechaza la hipótesis nula, pues el valor p es mucho menor que el nivel de significancia del 5%. Lo anterior implica seleccionar el modelo de efectos aleatorios.

```

. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

gini[pais,t] = Xb + u[pais] + e[pais,t]

Estimated results:

```

	Var	sd = sqrt(Var)
gini	87.54137	9.356355
e	15.68873	3.9609
u	49.04838	7.003455

```

Test: Var(u) = 0
      chibar2(01) = 295.23
      Prob > chibar2 = 0.0000

```

Figura 8.4: Prueba de Breusch-Pagan.

8.4.3 Modelo de panel de mínimos cuadrados con variable dicótoma de efectos fijos

Para determinar si existe o no heterogeneidad entre las unidades o empresas, la cual no es capturada por el modelo de coeficientes constantes, se plantea y se estima el modelo descrito en la [ecuación \(8.10\)](#).

$$\begin{aligned}
 lrent_{it} = & \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \cdots + \alpha_{39} D_{39i} \\
 & + \beta_2(lpop_{it}) + \beta_3(lavginc_{it}) + \beta_4 pctstu_{it} + U_{it}
 \end{aligned} \tag{8.10}$$

Para estimar el modelo anterior, se utiliza el comando:

Comando

```
regress gini yschooling poburb l.pais
```

```
. regress gini yschooling poburb i.pais
```

Source	SS	df	MS	Number of obs	=	234
Model	17369.2158	40	434.230396	F(40, 193)	=	27.68
Residual	3027.92437	193	15.6887273	Prob > F	=	0.0000
Total	20397.1402	233	87.5413742	R-squared	=	0.8516
				Adj R-squared	=	0.8208
				Root MSE	=	3.9609
gini	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yschooling	-.7200503	.4509988	-1.60	0.112	-1.60957	.1694689
poburb	.1536801	.0900458	1.71	0.089	-.02392	.3312802
país						
2	19.8154	3.134581	6.32	0.000	13.63297	25.99784
3	-12.62681	2.728392	-4.63	0.000	-18.00811	-7.245521
4	-3.988799	2.439249	-1.64	0.104	-8.799807	.8222095
5	13.41429	2.912234	4.61	0.000	7.670404	19.15818
6	11.18161	3.378275	3.31	0.001	4.518529	17.84469
7	-12.50201	2.917212	-4.29	0.000	-18.25572	-6.748303
8	-7.373937	2.326433	-3.17	0.002	-11.96244	-2.78544
9	-6.034165	2.853796	-2.11	0.036	-11.6628	-.4055324
10	-.8923855	2.547297	-0.35	0.726	-5.9165	4.131729
11	-6.817987	2.331761	-2.92	0.004	-11.41699	-2.218981
12	1.62811	2.764896	0.59	0.557	-3.825181	7.081401
13	-8.4597	3.044262	-2.78	0.006	-14.46399	-2.455405
14	-1.843247	4.462192	-0.41	0.680	-10.64417	6.957676
15	.4386923	4.382963	0.10	0.920	-8.205965	9.08335

Figura 8.5: Modelo panel de efectos fijos con variable dicótoma.

Los resultados de la estimación del modelo panel de efectos fijos con variable dicótoma ([Figura 8.5](#)) muestran que el coeficiente de la variable explicatoria yschooling es no significativo a un nivel del 5%, mientras que el coeficiente de la variable poburb es significativo a un nivel del 9%. Para determinar si existe heterogeneidad o algún efecto diferencial en el intercepto entre los 39 países, se plantean las hipótesis descritas en las [ecuaciones \(8.11\)](#) y [\(8.12\)](#).

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \dots = \alpha_{39} = 0 \quad (8.11)$$

$$H_1 : \text{Al menos un } \alpha_i \neq 0 \quad (8.12)$$

El estadístico utilizado para probar la hipótesis anterior es un F , el cual viene dado por la [ecuación \(8.13\)](#), como ya se ha presentado en secciones anteriores.

$$F = \frac{(SCR_0 - SCR_a)/m}{SRC_a/(n - k)}, \quad (8.13)$$

donde:

- SCR_0 = suma de cuadrados residuales bajo la hipótesis nula o restringida.
- SCR_a = suma de cuadrados residuales bajo la hipótesis alterna o no restringida.
- m = número de restricciones o el número de coeficientes de regresión que desaparecen del modelo original al imponer la restricción.

Si se rechaza la hipótesis nula, el modelo adecuado es el de efecto fijos. El **F** calculado con base en la suma de cuadrados de los residuos restringidos y no restringidos se muestra en la siguiente expresión:

$$F_c = \frac{(14812,92 - 3027,92)/38}{3027,92/(234 - 41)} = 19,77 \quad VpF = 0,00000$$

Por lo tanto, el modelo de panel con datos agrupados o de coeficientes constantes no es válido y se acepta el modelo de efectos fijos.

Una forma alterna de probar automáticamente en Stata la hipótesis anterior es siguiendo la ruta presentada para efectos fijos. El comando para estimar el modelo de efectos fijos es:

Comando

```
regress gini yschooling poburb, fe
```

```
. xtset pais year, yearly delta(5)
      panel variable: pais (strongly balanced)
      time variable: year, 1970 to 1995
      delta: 5 years

. xtreg gini yschooling poburb, fe

Fixed-effects (within) regression                               Number of obs     =      234
Group variable: pais                                         Number of groups  =       39

R-sq:
    within  =  0.0160                                         Obs per group:
    between =  0.0570                                         min  =         6
    overall =  0.0507                                         avg  =      6.0
                                                       max  =         6

                                                F(2, 193)      =     1.57
corr(u_i, Xb)  = -0.0526                                         Prob > F      =  0.2106


```

gini	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
yschooling	-.7200503	.4509988	-1.60	0.112	-1.60957 .1694689
poburb	.1536801	.0900458	1.71	0.089	-.02392 .3312802
_cons	31.55789	3.978178	7.93	0.000	23.71161 39.40418
sigma_u	8.4758174				
sigma_e	3.9608998				
rho	.82075804	(fraction of variance due to u_i)			

F test that all u_i=0: F(38, 193) = 19.77

Prob > F = 0.0000

Figura 8.6: Modelo de efectos fijos automático.

Como se observa en la [Figura 8.6](#), la salida no presenta las variables dummy, pero en la parte inferior el estadístico **F** y el valor **P** asociado coinciden con el calculado anteriormente y, por ende, se toma la misma decisión; es decir, que el modelo adecuado es el de efectos fijos por existir al menos un efecto diferencial distinto de 0.

8.4.4 Test de Hausman para escoger entre efectos fijos y aleatorios

Las pruebas anteriores determinaron que tanto el modelo de efectos fijos como el de efectos aleatorios son más adecuados que el de coeficientes constantes; ahora se debe escoger entre el modelo de efectos fijos y el de efectos aleatorios, para ello se recurre al test de Hausman, el cual plantea como hipótesis nula que los efectos individuales no están correlacionados con las variables explicatorias, y como hipótesis alterna que los efectos individuales están correlacionados con las variables explicatorias. Es decir, rechazar hipótesis nula es seleccionar el modelo de efectos fijos, mientras que no rechazarla es seleccionar el modelo de efectos aleatorios. La selección del modelo adecuado es importante, ya que:

1. Si los efectos no están correlacionados con las variables explicatorias, los estimadores obtenidos por el modelo de efectos aleatorios son consistentes y eficientes, pero los estimadores de efectos fijos son consistentes pero ineficientes.
2. Si los efectos están correlacionados con las variables explicatorias, los estimadores obtenidos por el modelo de efectos aleatorios son inconsistentes, pero los estimadores de efectos fijos son consistentes y eficientes.

Una forma alterna de plantear las hipótesis asociadas a la prueba de Hausman es:

$$H_0 : \text{EF y EA no difieren considerablemente}$$

$$H_1 : \text{Efectos fijos}$$

Para la realización de la prueba de Hausman, que permite determinar cuál modelo de datos de panel es el más adecuado, se siguen estos pasos:

1. Correr el modelo mediante efectos fijos.

Comando

```
regress gini yschooling poburb, fe
```

2. Guardar las estimaciones del modelo de efectos fijos y asignarles un nombre, en nuestro caso, fijos. Usar el comando:

Comando

```
estimate store fijos
```

3. Correr el modelo mediante efectos aleatorios.

Comando

```
regress gini yschooling poburb, re
```

4. Guardar las estimaciones del modelo de efectos aleatorios y asignarles un nombre, en nuestro caso, aleatorios. Usar el comando:

Comando

```
estimate store aleatorios
```

5. Realizar el test de Hausman a través del comando:

Comando

```
hausman fijos aleatorios
```

```
. hausman fijos aleatorios
```

	Coefficients			
	(b) fijos	(B) aleatorios	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
yschooling	-.7200503	-1.135576	.4155253	.2845482
poburb	.1536801	.1232798	.0304003	.0702286

b = consistent under Ho and Ha; obtained from xtreg

B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

```
chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)  
= 13.69  
Prob>chi2 = 0.0011
```

Figura 8.7: Prueba de Hausman.

La Figura 8.7 muestra los resultados de la prueba de Hausman, en la cual se rechaza la hipótesis nula, pues el valor chi cuadrado calculado para 2 gl (grados de libertad) es 13.69 con un valor p= 0.0011, esto significa que rechazamos el modelo de efectos aleatorios a favor del de efectos fijos.

A continuación se presenta el archivo .do utilizado en este capítulo, vale destacar que la única modificación para otra aplicación sería tener en cuenta la frecuencia de los datos que se utilizarán. Si los comandos **xtserial** y **xttest0** no se encuentran instalados, debe hacerlo con los comandos **findit xtserial** y **findit xttest0**.

8.5 Apéndice

Apéndice 1.

Archivo do

```
* Ubicar primero su base de datos  
use "J:gini.dta", clear  
* Declarar el panel teniendo en cuenta que hay observaciones  
* cada 5 años  
xtset pais year, yearly delta(5)  
* Estimar el modelo pool con coeficientes constantes  
regress gini yschooling poburb  
*Verificar si existe correlación en el error idiosincrático  
xtserial gini yschooling poburb, output  
*Estimar efectos aleatorios  
xtreg gini yschooling poburb, re  
*Verificar si existe autocorrelación en el modelo de efectos aleatorios  
* a través de la prueba de Breusch-Pagan  
xttest0  
* Si rechaza, es más adecuado efectos aleatorios.  
*Estimar efectos fijos con variables dummy y determine a través de  
*mínimos cuadrados restringidos, si existe heterogeneidad entre las  
*unidades  
regress gini yschooling poburb l.pais  
* Calcular el F de mínimos cuadrados restringidos  
*Stata calcula el F de mínimos cuadrados automáticamente  
xtreg gini yschooling poburb, fe  
*Determinar qué modelo es más adecuado, fijos o aleatorios, usando
```

*Hausman, estimar efectos fijos y guardar las estimaciones con un nombre,

*Repetir con efectos aleatorios

xtreg gini yschooling poburb, fe

estimate store fijo

xtreg gini yschooling poburb, re

estimate store aleatorio

hausman fijo aleatorio

Referencias

- Cameron, A. C., y Trivedi, P. K. (2005). *Microeometrics: methods and applications*. Nueva York: Cambridge University Press.
- Cameron, A. C., y Trivedi, P. K. (2009). *Microeometrics using stata* (Vol. 5). Stata press College Station, TX.
- De Fanelli, A. M. G. (1989). Patrones de desigualdad social en la sociedad moderna: una revisión de la literatura sobre discriminación ocupacional y salarial por género. *Desarrollo económico*, 239–264.
- Enders, W. (2015). *Applied econometric time series*. John Wiley & Sons.
- Engle, R. F., y Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.
- Gujarati, D., y Porter, D. (2010). *Econometría (quinta edición)*. México: McGraw-Hill/Interamericana Editores, SS DE CV.
- Heckman, J. J. (1977). *Sample selection bias as a specification error (with an application to the estimation of labor supply functions)*. Cambridge, Mass., USA: Bureau of Economic Research.
- Johansen, S., y Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration—with applications to the demand for money. *Oxford Bulletin of Economics and statistics*, 52 (2), 169–210.
- Peñas, I. L. (2002). La discriminación salarial por razones de género: un análisis empírico del sector privado en España. *Reis*, 171–196.
- StataCorp. (2013). *Stata statistical software: Release 13*. College Station, TX: StataCorp LP.
- Wooldridge, J. M. (2010). *Introducción a la econometría: un enfoque moderno 4 ed.* Cengage Learning.