HERRAMIENTAS DE SOFTWARE APLICADAS AL MÉTODO DE REGRESIÓN LINEAL

Maria Paula Contreras Navarrete[1]
diciembre de 2011

Resumen:

El método de regresión lineal es una práctica estadística ampliamente utilizada para analizar la relación entre variables, teniendo gran variedad de aplicaciones en las diversas áreas económicas, políticas y sociales. Actualmente, muchos software econométricos y estadísticos han sido desarrollados para agilizar y favorecer el proceso de análisis y manejo de los datos, brindando cada vez más herramientas novedosas y haciendo que la aplicación de los métodos por parte del investigador se base principalmente en la interpretación de resultados.

En el presente documento se realizará una revisión sobre las principales características y herramientas brindadas por una serie de software en lo referente al desarrollo del método de regresión lineal, la comprobación de supuestos y su aplicación; con el objetivo de proveer al público una visión más amplia de la multiplicidad de instrumentos a su disposición.

Palabras Clave: método de regresión lineal, variables, software econométricos y estadísticos

SOFTWARE TOOLS APPLIED TO THE LINEAR REGRESSION METHOD

Abstract:

The method of linear regression is a statistical practice widely used to analyze the relationship between variables, having a vast amount of applications in various

[1] Estudiante de Economía de la Facultad de Ciencias Económicas de la Universidad Nacional de Colombia, y monitor de la Unidad de Informática y Comunicaciones de la Facultad de Ciencias Económicas. Correo Electrónico:

economic, political and social areas. Currently, many econometric and statistical software have been developed to expedite and smooth the progress of the process of analyzing and managing data, providing increasingly innovative tools and making that the application of methods by researchers is mainly based on the interpretation they give of results.

In this paper, the main features and tools offered by a number of software will be reviewed in terms of the appliance of linear regression, checking assumptions and their implementation in order to provide the public a wider vision of the multiplicity of instruments at their disposal.

Keywords: method of linear regression, variables, econometric and statistical software

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE CIENCIAS ECONÓMICAS
UNIDAD DE INFORMÁTICA Y COMUNICACIONES

Director Unidad Informática:
Henry Martínez Sarmiento

Tutor Investigación:
Juan Carlos Tarapuez Roa

Coordinadores:
Jasmin Guerra Cárdenas
Juan Felipe Reyes Rodríguez

Coordinador Servicios Web:
John Jairo Vargas

Analista de Infraestructura y Comunicaciones:
Diego Alejandro Jiménez Arévalo

Analista de Sistemas de Información:
Víctor Hugo Ramos Ramos

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE CIENCIAS ECONÓMICAS
UNIDAD DE INFORMÁTICA Y COMUNICACIONES

Estudiantes Auxiliares:

Camilo Alexandry Peña Talero
Cristian Andrés Hernández Caro
Claudia Patricia Ospina Aldana
Daniel Francisco Rojas Martín
David Camilo Sánchez Zambrano
David Mauricio Mahecha Salas
Diego Esteban Eslava Avendaño
Edward F. Yanquen Briñez
Gloria Stella Barrera Ardila
Iván Albeiro Cabezas Martínez
Javier Alejandro Ortiz Varela
Jeimmy Paola Muñoz
Juan Carlos Tarapuez Roa
Juan David Vega Baquero
Juan Fernando López Prieto
Leonardo Alexander Cárdenas
Leidy Esther Fernández Coba
Lina Marcela Igua Torres
María Paula Contreras Navarrete
Paola Alejandra Alvarado Castillo
Viviana Contreras Moreno
Viviana María Oquendo

Este documento es resultado de un trabajo conjunto y coordinado de los integrantes de la Unidad de Informática y Comunicaciones de la Facultad de Ciencias Económicas de la Universidad Nacional de Colombia.

HERRAMIENTAS DE SOFTWARE APLICADAS AL MÉTODO DE REGRESIÓN LINEAL

Contenido

1. INTRODUCCIÓN

En economía y cualquier otra disciplina académica, la econometría y especialmente la estadística constituyen una parte fundamental del análisis de los diferentes fenómenos sociales. A partir de ahí, la econometría toma gran variedad de esas herramientas estadísticas para evaluar modelos y metodologías que fundamenten y reafirmen la pertinencia de la teoría en la realidad.

En este sentido, el método de regresión lineal es una técnica que evalúa con precisión la existencia de relaciones entre ciertas variables y su utilización se ha extendido a muchos campos económicos, sociales y políticos, en donde los teoremas necesitan una herramienta de medición que permite revisar, analizar e interpretar sus aplicaciones en el mundo real.

Por otro lado, durante la formación académica de los estudiantes la aplicación práctica en software queda limitada a la disposición de las clases que se encargan de aplicarlos, proveyendo estas únicamente los códigos necesarios para realizar las actividades correspondientes. Se evidencia que la mayoría de los estudiantes no adquieren conocimiento básico de todas las herramientas disponibles para aplicar lo aprendido teóricamente, tomando una visión sesgada y quedando en una posición de desventaja al enfrentarse al mundo real como profesionales.

2. CONTENIDO

2.1. Métodos Estadísticos

En economía, todo análisis debe estar fundamentado en métodos estadísticos que disminuyan la brecha entre la teoría y la práctica, validando de forma consistente los teoremas abordados en la academia y permitiendo un análisis e interpretación más profundos y completos de las situaciones que se presentan a diario. Así, la estadística se convierte en una herramienta esencial para el desarrollo integral de cualquier proceso de observación, exploración o investigación a través del cual se busca la obtención de resultados confiables que permitan llegar a conclusiones importantes sobre el comportamiento de los diferentes fenómenos sociales.

Como toda disciplina, la estadística está dividida en ramas que permiten abordar con más especificidad cierto tipo de procesos, reconociéndose especialmente dos: la estadística descriptiva, la cual se encarga exclusivamente de la recolección y presentación de los datos,

y la inferencial, la cual a partir de datos muestrales[2] realiza estimaciones y generalizaciones sobre una cantidad mayor de datos.

2.1.1. Estadística Descriptiva

Antes de profundizar en el planteamiento y aplicación del método de regresión lineal, es fundamental familiarizarse con algunos conceptos estadísticos que facilitan el entendimiento del proceso. En otros términos, si lo que nos interesa en el análisis es determinar si existe algún tipo de relación entre dos o más variables, es primordial que primero conozcamos y sepamos interpretar en detalle los datos que las representan para así poder concluir sobre el comportamiento de dichas variables y las implicaciones que este puede tener sobre otras.

Es aquí donde la estadística descriptiva entra a jugar un papel importante. Ésta no solo se ocupa de recolectar los datos, sino también de organizarlos, tratarlos, resumirlos y presentarlos al investigador de una manera precisa, a través de tablas y gráficos, posibilitando así su manejo e interpretación. Igualmente permite realizar el cálculo de ciertos parámetros que recogen información y características importantes del comportamiento de los datos y que pueden reunirse en cuatro grandes categorías de acuerdo a lo que describen: posición, centralización, dispersión y forma. A continuación, se hará una breve revisión sobre cada uno de ellos.

2.1.1.1. Medidas de Posición

Estas dividen un conjunto ordenado de datos en intervalos que contienen el mismo número de elementos, simplificando la tarea de ubicar algún elemento específico dentro de un gran conjunto de datos. La medida más utilizada en este campo se conoce como Percentil, el cual, para una variable discreta "se define el percentil de orden k, como la observación, P_k, que deja por debajo de sí el k % de la muestra"[3].

Otra medida la cual es un caso particular de los percentiles se conoce como Cuartil y divide el conjunto de datos en únicamente cuatro intervalos de igual tamaño. Están organizados de la siguiente manera:

- Primer cuartil o cuartil inferior (Q_1): el 25% de los datos se encuentran por debajo de él, es decir son menores a su valor.

[2] Obtenidos a partir de una muestra.

[3] Tomado el 26 de Agosto de 2011 de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema3.pdf. Página 7.

- Segundo cuartil o cuartil intermedio (Q_2): el 50% de los datos se encuentran por debajo de él, es decir que su valor es mayor a la mitad de los datos. Coincide con la mediana.

- Tercer cuartil o cuartil superior (Q_3): El 75% de los datos se encuentran por debajo de su valor.

Finalmente se encuentran los Deciles, los cuales dividen el conjunto de datos en 10 intervalos iguales cada uno con una coincidencia respectiva de los percentiles 10, 20, 30,..., 90.

2.1.1.2. Medidas de centralización

Consisten en medidas que señalan valores sobre los cuales los datos de la muestra se encuentran centrados y alrededor de los cuales se agrupan, siendo a su vez útiles para descubrir la existencia de datos con comportamiento atípico.

La Media o promedio aritmético es la medida de centralización más utilizada debido a que expresa la concentración de los datos en términos de todos los elementos de la muestra. Se calcula a partir del cociente entre la sumatoria de todas las observaciones y el número de ellas.

Por otro lado se encuentra la Mediana la cual representa la mitad de la sucesión del conjunto de datos ordenados, indicando así que el 50% de los valores de la muestra es menor a esta y el otro 50% es mayor.

Adicionalmente la centralización se representa por medio de la Moda, la cual indica la mayor frecuencia con la que aparece un dato, es decir el valor del dato que más veces se repite.

2.1.1.3. Medidas de dispersión

Estas medidas son tomadas en referencia con base en las medidas de centralización e indican la mayor o menor concentración que los datos tengan respecto a estas (sus valores). Al medir la variabilidad de los valores respecto al valor que fue determinado como central, son una prueba de si las medidas de centralización están realmente representando a la información en su conjunto. Dentro de la categoría se destacan tres medidas:

- Rango muestral: representa la diferencia entre el valor máximo de las observaciones y el mínimo. Ésta medida sin embargo posee ciertos problemas

dentro de su cálculo dentro de los cuales se destacan: al ser la diferencia entre dos dígitos no utiliza todas las observaciones de la muestra y se puede afectar por algún valor muy extremo.

- Varianza (s^2): es la sumatoria de los desvíos de la media (distancias entre cada dato y la media) al cuadrado sobre el número de observaciones de la muestra. Al estar elevada al cuadrado, no tiene las mismas unidades que las demás variables, lo que no me permite compararlas entre sí.
- Desviación Estándar (s): Esta definida como la raíz cuadrada positiva de la Varianza, solucionando así el problema de trabajar con una medida que se encuentra en unidades diferentes a las de las demás variables.

2.1.1.4. Medidas de Forma

En cuanto a la forma en la cual se distribuyen los datos, las principales medidas a determinar son la Simetría y la Kurtosis. En este caso el interés se centra primero en analizar si los datos se distribuyen simétricamente respecto a alguna medida de centralización, o si presentan algún sesgo (están más concentrados) hacia la derecha o izquierda. Una vez indicada la simetría o asimetría de los datos, es necesario saber, por medio de la Kurtosis, si la curva es apuntada o relativamente plana (esta es una medida que está directamente relacionada con la concentración de los datos hacia la moda).

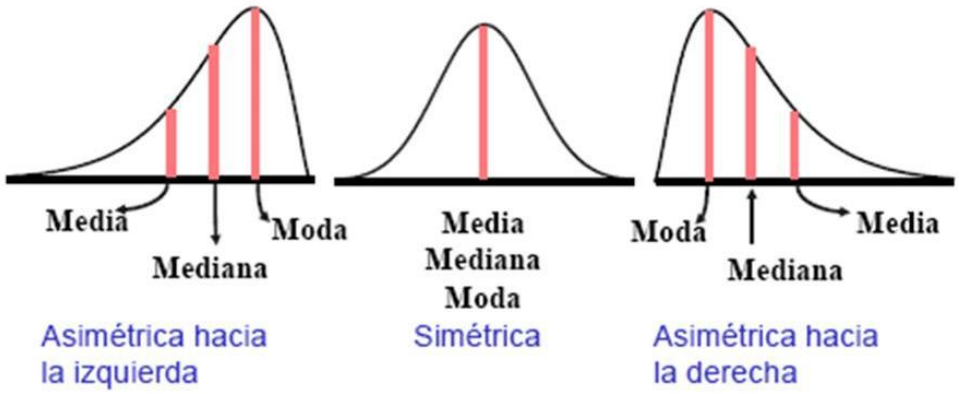


Ilustración 1

Nota: Para la interpretación de ambas medidas se toma como referencia la conocida Distribución Normal.

2.2. Software Econométrico y Estadístico

En la actualidad, existe multiplicidad de software que reúnen un sinnúmero de características y funcionalidades que brindan a los usuarios una mayor simplicidad y comodidad a la hora de realizar el cálculo y la estimación desde estadísticas básicas como las mencionadas anteriormente, hasta la aplicación más rigurosa de diferentes técnicas o métodos para la construcción de modelos econométricos.

Nuestro centro de interés es observar las diferentes herramientas brindadas por un conjunto determinado de software para la aplicación del método de regresión lineal. Así, en la presente sección se pretende presentar y evaluar las alternativas ofrecidas, en términos de la obtención de las medidas de estadística descriptiva, desde el entorno de los software R-project, RATS, Stata, SPSS y SAS. Esto con el fin de tener una visión más amplia de todas las herramientas que se encuentran disponibles en el mercado y a su vez hacer un paralelo entre cada una de ellas.

2.2.1. Stata 11.0

Stata es un software estadístico completo e integral, el cual proporciona todo lo necesario para el análisis y la gestión de datos.[4] Su suite es muy interactiva y se caracteriza por tener una interfaz dinámica tanto en términos de variedad y utilidad de los menús desplegables como en una línea de código intuitiva. Sata maneja archivos de extensión .dta.[5]

La interfaz está dividida en 5 partes principales: La cinta de opciones; la ventana de Comandos, en la cual aparece la lista de instrucciones elaboradas; la ventana de variables, en la cual se despliegan todas las variables que contenga la base de datos; la ventana de resultados, la central; y la ventana de códigos, o el espacio en donde se van a digitar las ordenes.

[4] Tomado el 26 de Agosto de 2011 de http://www.stata.com/whystata/

[5] Es importante resaltar que si desea trabajar con un archivo de Excel es necesario que este guardado en formato Texto(delimitado por tabulaciones) y que el nombre bajo el cual se guarda no puede contener espacios.

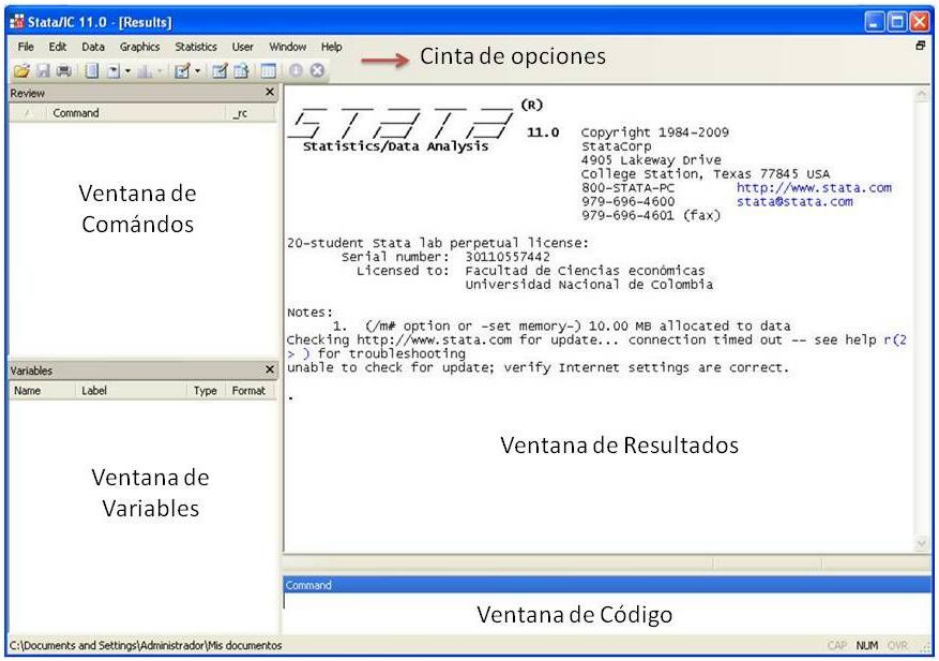


Ilustración 2. Interfaz de Stata

Stata permite trabajar de igual forma por medio de un Script, en el cual se digitan instrucciones sin ser ejecutadas de forma inmediata. Para abrir un nuevo Script debe dirigirse a la pestaña ***Window*** en la cinta de opciones y seleccionar la última opción: ***Do-file Editor***[6]. En el momento en que se desee ejecutar los comandos, puede: seleccionar el ícono  el cual hará que se ejecute todo o seleccionar el comando que desea y presionar Ctrl+d.

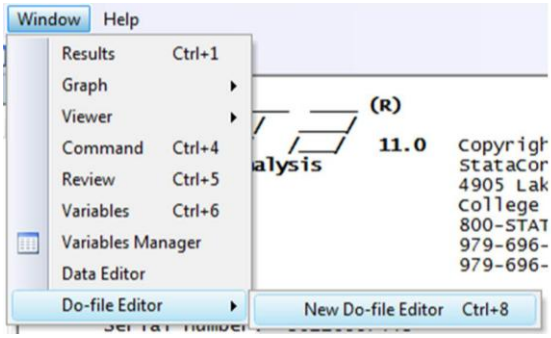


Ilustración 3. Abrir un Script en Stata

[6] En el editor es posible crear un archivo .log en el cual se va a guardar todo lo que se realice. Por medio del siguiente código *log using "nombre del archivo".log, replace*. Cuando termine de trabajar deberá escribir *log close* para completar el proceso.

Para cargar los datos se debe utilizar el comando ***insheet using "nombre del archivo".txt.*** Para visualizar los datos ya cargados debe dirigirse en la ventana principal de Stata a la pestaña ***Window>Data Editor.***

Para introducir al usuario al funcionamiento del software, se va a trabajar sobre una base de datos de ejemplo que contiene Stata. Para abrirla debe dirigirse a ***File***(Archivo) → ***Example Datasets*** (bases de datos de ejemplo), seleccionar las que están instaladas de forma predeterminada (*Example Datasets installed with Stata*) y finalmente seleccionar la opción que le permitirá utilizarla, *use*.

Inmediatamente aparecerá en la ventana de resultados un comando que le indica que la base de datos fue cargada exitosamente. Igualmente el contenido de la base de datos aparecerá en la ventana de variables. Igualmente se hubiera podido cargar la base de datos por medio de la ventana de códigos: para esto deberá escribir el comando **sysuse** seguido del nombre del archivo que desea abrir, en este caso **auto.dta** y presionando *Enter*.

```
. sysuse auto.dta
(1978 Automobile Data)
```

Ilustración 4. Cargar base de datos

En muchos casos es de suma importancia para los investigadores conocer información más detallada sobre el contenido de la base de datos y especialmente en la econometría las estadísticas básicas sobre las variables son necesarias para realizar el análisis. Para esto, Stata tiene en la cinta de opciones un ícono llamado ***Statistics,*** el cual además de contener la opción para visualizar las estadísticas básicas de las variables, va a permitir más adelante la aplicación de las diferentes metodologías (regresión lineal, métodos de análisis de series de tiempo, modelos no lineales, etc.).

La obtención de las estadísticas básicas puede ser de dos formas:

1. Seleccionando en orden los íconos ***Statistics >Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics*** y finalmente validando la instrucción a través del botón OK.
2. Escribiendo en la ventana de códigos el comando **summarize** y presionando *Enter*.

Independientemente de la forma en que se realice, en la ventana de resultados aparecerá un cuadro que muestra: 1. El número de observaciones, 2. La media, 3. La desviación estándar, 4. El mínimo y 5. El máximo de cada variable; como se muestra en la ilustración 5.

. summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

Ilustración 5. Estadística descriptiva

Esta es una vista general de las estadísticas de todas las variables, sin embargo y para más exactitud, Stata ofrece otro comando para analizar cada variable por separado, permitiendo que ningún detalle se escape. Este comando recibe el nombre de **codebook** y puede ser escrito así en la ventana de códigos u obtenerse el mismo resultado a través de la cinta de opciones: *Data > Describe data > Describe data contents (codebook);* inmediatamente aparecerá una ventana que le indicará si desea realizar la operación para una sola variable, caso para el cual deberá poner en el espacio *Variables* el nombre de la variable; o para todas las variables, caso en el cual deberá dejar el espacio vacío.

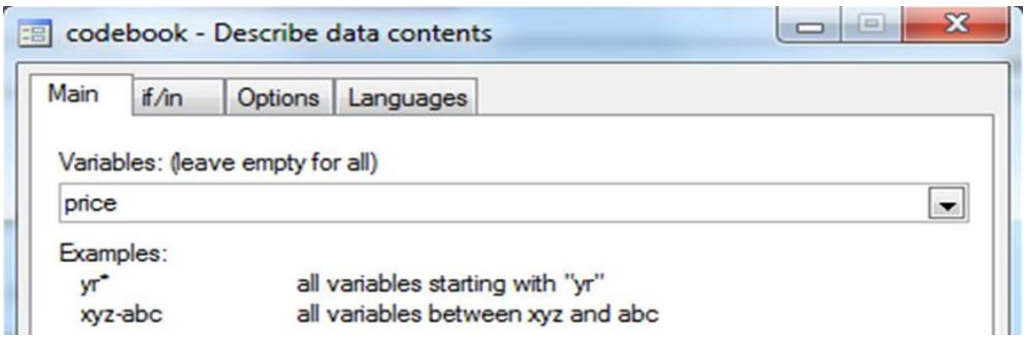


Ilustración 6.Codebook

En la ventana de resultados aparecerá:

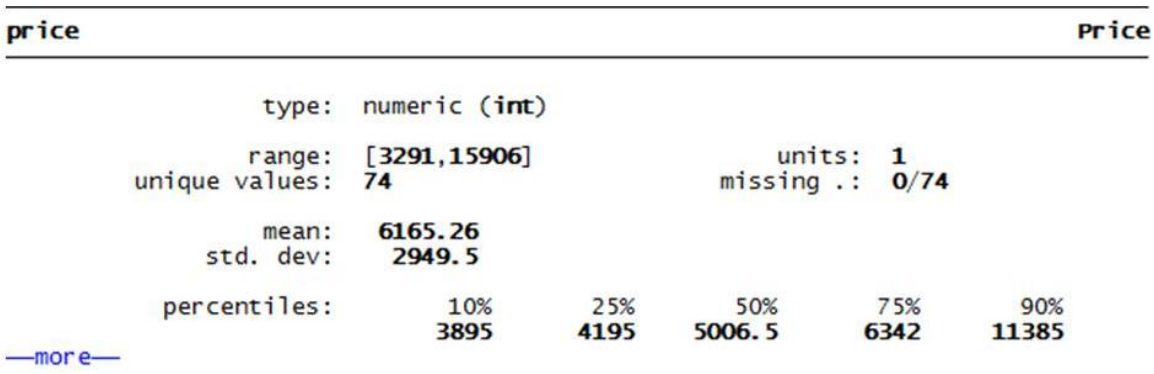
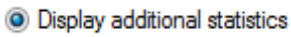


Ilustración 7. Output

Nota: La palabra en azul *more* indica que más información se encuentra disponible. Para visualizarla solo debe seleccionar la palabra.

A partir del comando que resume las estadísticas básicas de las variables podemos obtener más información esencial. Escribiendo el comando **summarize "el nombre de la variable", detail** (en este caso summarize price, detail) o bien volviendo al menú *Statistics >Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics* y eligiendo en la ventana que aparece la opción



Esta acción mostrará en más detalle los percentiles, la varianza de la variable, su simetría y kurtosis.

2.2.2. R-Project

R es un lenguaje y entorno (sistema) que provee gran variedad de técnicas estadísticas y gráficas para la aplicación de modelos lineales o no lineales, la realización de pruebas estadísticas básicas, análisis de series de tiempo, entre muchos otros. R se encuentra disponible como software libre[7] y es altamente extensible, ofreciendo la posibilidad de incluir cuando sea necesario nuevos paquetes desarrollados por la comunidad que satisfagan la necesidad del usuario.

[7] Bajo la licencia GNU o General Public License

La interfaz de R es muy sencilla: en la parte superior se encuentra ubicada la cinta de herramientas y algunos botones de uso común tales como abrir, guardar, copiar, etc. La demás parte está compuesta por la consola principal, en la cual todos los comandos van a ser ejecutados (presionando *Enter*) y los resultados visualizados. Adicionalmente, a medida que se vaya digitando el código, algunas ventanas complementarias irán apareciendo.

De manera alterna, R permite trabajar en un editor o script en donde se digitan los comandos pero no se ejecutan inmediatamente, lo cual brinda al usuario mayor comodidad. Gran cantidad de códigos pueden ser copiados pero únicamente se van a ejecutar en la consola presionando la tecla F5.

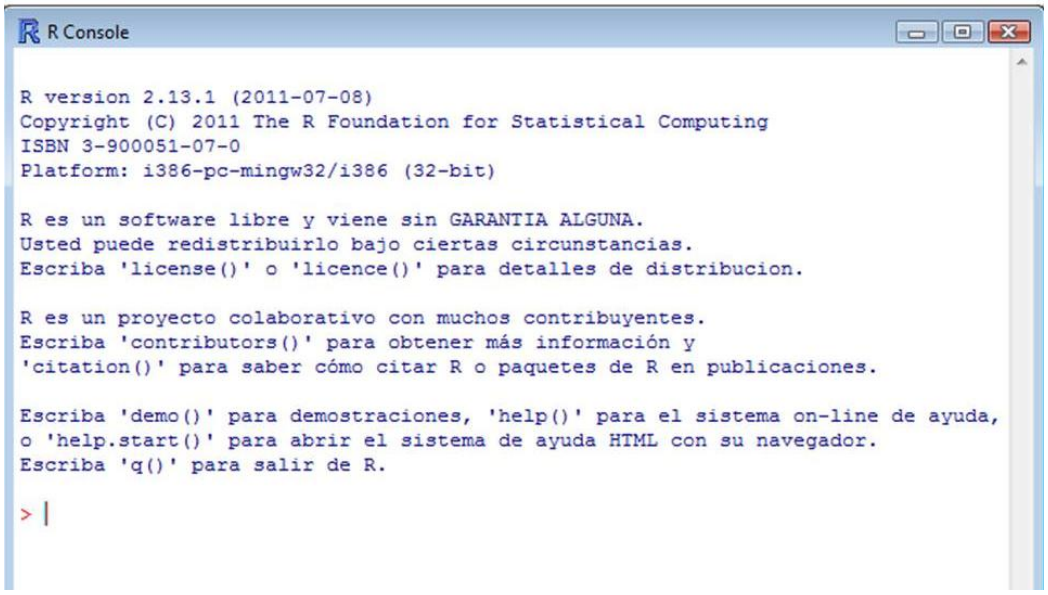


Ilustración 8. Consola principal de R-Project

Para comenzar a utilizar el software debe cargarse inicialmente una base de datos; para este efecto, es importante mencionar que en R se puede trabajar con varios tipos de archivo: .csv,.txt y .xls. Esta es una gran facilidad porque permite que el usuario maneje sus bases de datos en Excel y luego las importe para trabajar directamente en ellas. En este caso es necesario que en Excel el archivo quede guardado bajo el formato .csv (delimitado por comas).

Para comenzar a utilizar el programa y cargar las bases de datos sin problema alguno, es necesario dirigirse a la opción *Archivo* > *Cambiar dir…*, y escoger el destino en el cual se encuentran localizados los archivos sobre los cuales se va a trabajar, esto con el fin de que R los encuentre rápidamente. Igualmente antes de comenzar a insertar las órdenes, es recomendable que se limpie la memoria del software para evitar posibles incongruencias;

para esto se utiliza el comando **rm(list=ls())**. Si el usuario necesita conocer más información acerca de las funciones de un comando, podrá digitar **help()**, metiendo dentro del paréntesis el código.

Para cargar los archivos, el código que debe transcribirse es **read.csv2.** La formación del código comienza por el nombre que se le asigna al nuevo objeto formado (esto con el fin de que el software lo identifique fácilmente), seguido de la instrucción formal *read.csv2* y un paréntesis en donde debe especificarse el nombre del archivo original y si este tiene etiquetas o títulos. Luego de que el objeto ha sido creado, para visualizarlo es necesario llamarlo escribiendo nuevamente su nombre. Todo el proceso se enseña en la ilustración 9:

```
> Base1 <- read.csv2("Base1.csv", header=T)
> Base1
```

Ilustración 9

Nota: El símbolo <- representa la asignación de la orden al objeto Base1. Este procedimiento debe hacerse para todos los comandos debido a que R va a reconocer únicamente los objetos creados a los cuales se les asignó una instrucción. Puede ser sustituido por el símbolo =.

R permite transformar la forma de los datos haciendo que puedan ser leídos y entendidos como matrices, siguiendo el código **rh1=as.matrix(Base1)**[8] **.** Si desea visualizar el objeto en una ventana adicional y de forma más organizada debe escribir **View()**, metiendo entre paréntesis el nombre del objeto que desee. En este caso la instrucción View(rh1) genera el siguiente resultado:

[8] Dentro del paréntesis debe ir especificado el nombre del objeto con el cual se reconocen los datos.

Archivo

	CIUDADES	RH_1	RH_2	RH_3	RH_4	RH_5	RH_6	RH_7	RH_8	RH_9	RH_10	RH_11
1	Armenia	284120	0.00565	0.4640147	0.5373766	0.0880	1.095	0.266	22.4	0.354	3037.5	0.6282
2	Barranquilla	1163007	0.00721	0.4883377	0.5554151	0.0922	1.083	0.306	22.5	0.218	1054.1	0.8594
3	Bogotá, D.C.	7050228	0.01513	0.5631922	0.6293213	0.0973	0.999	0.538	22.6	0.447	1042.5	0.7052
4	Bucaramanga	520080	0.00344	0.5108233	0.5666209	0.0760	1.192	0.326	22.2	0.378	1068.0	1.0000
5	Cali	2169801	0.01163	0.5528726	0.6205193	0.0831	1.019	0.223	22.0	0.313	626.8	0.8590
6	Cartage	912674	0.01115	0.4839377	0.5469014	0.1213	1.207	0.147	24.4	0.215	1866.8	0.5235
7	Cúcuta	600049	0.01042	0.5193251	0.5768769	0.1133	1.147	0.217	22.4	0.181	1245.2	0.8040
8	Ibagué	509796	0.01130	0.5166996	0.5919223	0.1073	0.983	0.181	23.4	0.252	1677.9	0.4559
9	Manizales	383483	0.00460	0.4609772	0.5199243	0.0769	1.143	0.230	22.2	0.326	2233.2	0.7433
10	Medellín	2264776	0.01123	0.4963598	0.5592503	0.1256	1.219	0.300	26.3	0.310	1167.4	1.0000
11	Montería	390996	0.01562	0.4797224	0.5548727	0.1587	1.109	0.151	26.1	0.160	836.2	0.6002
12	Neiva	322098	0.00950	0.4980671	0.5510142	0.1131	1.100	0.179	23.3	0.282	1229.8	0.8032
13	Pasto	394074	0.01475	0.4974571	0.5674950	0.1115	1.052	0.128	21.8	0.397	807.1	1.0000
14	Pereira	448971	0.00607	0.4839616	0.5496611	0.1007	1.124	0.283	22.6	0.303	1557.0	0.6990
15	Popayán	261694	0.00805	0.5230494	0.5767888	0.0915	1.135	0.163	21.5	0.395	1994.4	0.6892
16	Riohacha	184847	0.04818	0.4279488	0.5072428	0.2713	1.033	0.152	26.1	0.188	1693.9	0.8537
17	Santa Marta	428374	0.01553	0.4724664	0.5337993	0.1244	1.005	0.116	22.9	0.228	1238.5	0.7396
18	Sincelejo	245180	0.01566	0.4713240	0.5370866	0.1612	1.228	0.107	24.9	0.159	948.7	0.6234
19	Tunja	161209	0.02256	0.5285888	0.5821245	0.1023	1.036	0.245	22.1	0.314	1221.9	0.9228
20	Valledupar	373872	0.02667	0.4363452	0.5069274	0.1587	1.017	0.147	24.5	0.258	757.3	0.9383
21	Villavicencio	400475	0.02595	0.5434616	0.6030498	0.1017	1.193	0.174	25.4	0.281	661.8	0.6042

Ilustración 10. Visualización de los objetos

Dado que no todas las variables de la matriz están en formato numérico, para la manipulación de los datos es mejor eliminar la primera columna que contiene el nombre de las ciudades. Para esto, utilice el comando **datosrh=Base1[,]**, en donde en la primera posición indica el número de filas y en la segunda el de columnas. Como desea eliminar una columna, el paréntesis debe contener los elementos de esta forma **[,-1].** En cuanto al tema central de la estadística descriptiva, R ofrece una serie de comandos simples y fáciles de recordar a través de los cuales podemos visualizar los principales estadísticos de medición y los cuales se especifican de la siguiente forma:

- Para visualizar los estadísticos más básicos como el mínimo, máximo, la mediana y la media de cada uno de los datos del conjunto, el comando será **summary(),** colocando entre paréntesis el nombre otorgado a la matriz sin la variable texto**.**

```
> estadisticas=summary(datosrh)
> estadisticas
      RH_1               RH_2               RH_3               RH_4
 Min.   : 161209   Min.   :0.00344   Min.   :0.4279   Min.   :0.5069
 1st Qu.: 322098   1st Qu.:0.00805   1st Qu.:0.4725   1st Qu.:0.5374
 Median : 400475   Median :0.01130   Median :0.4964   Median :0.5554
 Mean   : 927134   Mean   :0.01430   Mean   :0.4961   Mean   :0.5607
 3rd Qu.: 600049   3rd Qu.:0.01562   3rd Qu.:0.5193   3rd Qu.:0.5769
 Max.   :7050228   Max.   :0.04818   Max.   :0.5632   Max.   :0.6293
```

Ilustración 11. Estadística descriptiva en R

- En cuanto a las medidas de dispersión, los códigos corresponden a las primeras partes de sus nombres, es decir **var() y sd(),** representando respectivamente la varianza y la desviación estándar (sd responde en este caso a las siglas en inglés de Standard Deviation). Dentro del paréntesis se especifica el objeto sobre el cual se va a hacer el estudio descriptivo.

- Como se había mencionado anteriormente, la estadística descriptiva también brinda grandes herramientas de análisis en torno a la presentación de los datos por medio de gráficos. R es una excelente herramienta para la elaboración de gráficos dentro de los cuales se destacan los boxplots, histogramas y los gráficos de dispersión, los cuales pueden llegar a ser una herramienta muy útil a la hora de compararlos con las demás medidas explicadas.

    - Los códigos correspondientes son: **boxplot(), hist() y pairs().** Dentro del argumento de cada comando es posible modificar las opciones y estilos de cada gráfico; para obtener más información podrá introducir **?boxplot** (lo que es equivalente a help(boxplot)) o el nombre del gráfico que desee.

      A continuación se ilustra el gráfico de dispersión, el cual permite observar si hay una posible correlación entre las variables: luego de introducir el código, una ventana nueva aparecerá con el gráfico.

```
> pairs(datosrh, main="Gráfico de dispersión de los datos")
```

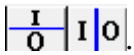
Ilustración 12. Gráfico de dispersión en R

Nota: Es importante resaltar que todas las operaciones de obtención de los estadísticos descriptivos se realizaron con la matriz que contenía únicamente variables numéricas. Si su base de datos contiene variables representadas por texto, es necesario que para la obtención de dichas medidas que representan el comportamiento de los datos, estas sean eliminadas.

2.2.3. WinRATS 7.2.

Comúnmente conocido como RATS (Regression Analysis of Time Series), es un reconocido paquete de software para análisis econométrico y de series de tiempo. Se caracteriza por ser rápido, eficiente, flexible y comprensivo[9] y es un software muy utilizado especialmente en universidades y corporaciones alrededor del mundo. Dentro de sus herramientas, RATS incluye la estimación de mínimos cuadrados lineales y no lineales, modelos ARIMA, GMM, ARCH, GARCH, entre otros, y es uno de los pocos software que ofrece las capacidades de análisis espectral.

RATS trabaja por defecto con archivos de extensión *.PRG* , aunque también recibe archivos de código fuente *.SRC* y archivos de texto. Adicionalmente permite leer información de Excel 2007, Stata y Eviews. Para esto, las hojas de cálculo en particular, deben tener especificada en la primera columna la periodicidad de las observaciones en formato fecha.

La interfaz del software se caracteriza por ser simple y amable al usuario: en la parte superior se encuentran las barras de herramientas y en la pantalla central se encuentra el área de trabajo la cual cuenta con una ventana de entradas (Input) y otra de salidas (Output), las cuales pueden organizarse de forma vertical u horizontal con la ayuda de los botones  .

Es importante resaltar la presencia de otros dos botones en la barra superior:  . El primero sirve para cambiar el modo de la ventana de entradas, pasarlo de Ready (listo para ejecutar) a Local Edit (editar); es de gran importancia tener esto en cuenta para no tener problemas cuando se deseen insertar nuevas órdenes sin necesidad de ejecutarse inmediatamente. Es importante mencionar que el modo de ejecutar las instrucciones es presionando la tecla "Enter". Y el último es para limpiar el programa cuando sea necesario, debido a la necesidad de ejecutar nuevas instrucciones

[9] Tomado de la página web del producto: http://www.estima.com/ratsmain.shtml

independientes a las que ya se habían insertado o para evitar que el programa presente inconsistencias.

Nota: Se va a trabajar con una base de datos que incluye las variables Inversión, PIB real, Tipo de Cambio real, Tasa de Inflación y Tipo de Interés real y cuyas observaciones son trimestrales desde el año 1994 hasta el 2007.

El primer paso a seguir es importar los datos e imprimirlos en el output, para lo cual se debe utilizar la serie de comandos de la ilustración 13,

```
call 1994 1 4
all 2007:04
open data
data(format=xls,org=obs)
print /
```

Ilustración 13. Importar datos

En donde ***call*** y ***all*** le indican al software llamar los datos desde una fecha específica hasta otra (los números que aparecen al lado de los años indican que la periodicidad de los datos es trimestral y que desea llamar los datos hasta el último trimestre del 2007[10]). ***Open data*** *y* ***data*** le ordenan al programa abrir los datos con un determinado formato desde una ubicación en el computador; cuando ejecute ambas instrucciones una ventana emergerá para localizar el archivo sobre el cual va a trabajar. Finalmente el comando ***print /*** le dice al programa que muestre o imprima los datos ya leídos (la barra / significa toda la información), los cuales aparecerán en la ventana de salidas.

Para reconocer las variables más fácilmente, RATS permite cambiarles el nombre para así identificar cual es la endógena y cuales las exógenas. Para esto se utiliza el comando **set** y[11] = "nombre de la variable original". En el caso del ejemplo, se haría de la siguiente manera:

```
set y = INVEREX
set x1 = PIBR
set x2 = E
   ⋮
```

Ilustración 14. Cambio de nombre a las variables

[10] En caso de ser datos anuales, no es necesario especificar lo último, pero si se desea puede colocarse un 1.

[11] En el caso de la variable endógena.

En RATS, las estadísticas básicas se obtienen de forma individual para cada variable, sin embargo también es posible obtener una caracterización general de los datos por medio del comando ***table***, el cual especifica las series, el número de observaciones, la media, desviación estándar, mínimo y máximo, tal como se enseña en la ilustración 15[12].

Series	Obs	Mean	Std Error	Minimum	Maximum
INVEREX	56	39652675,5770	1286568,8408	2130254,0000	7426559,1320
PIBR	56	19756862,0357	2296941,5015	16483795,0000	25839016,0000
E	56	112,5666	12,7545	89,0371	137,2577
INF	56	11,7298	6,8148	4,0327	23,6323
IR	56	13,6486	6,3365	7,0190	37,4807

Ilustración 15. Estadística descriptiva en RATS

Para la obtención de estadísticas más detalladas sobre las variables es necesario utilizar el comando ***statistics*** seguido de la variable deseada; así obtendrá lo siguiente:

```
statistics y
```

Statistics on Series Y			
Quarterly Data From 1994:01 To 2007:04			
Observations	56		
Sample Mean	3965267,557714	Variance	1655259382175,0400
Standard Error	1286568,840822	of Sample Mean	171924,993102
t-Statistic (Mean=0)	23,063939	Signif Level	0,000000
Skewness	0,867333	Signif Level (Sk=0)	0,009918
Kurtosis (excess)	0,464419	Signif Level (Ku=0)	0,505592
Jarque-Bera	7,524424	Signif Level (JB=0)	0,023232

Ilustración 16. Estadísticas adicionales

Adicionalmente, como en R-project se podrán observar el comportamiento gráfico de las variables de la siguiente forma

```
graph 1
# x2
```

Ilustración 17. Comando para graficar en RATS

[12] Los resultados en RATS aparecen de forma desordenada. Para su presentación es necesario recurrir a otras herramientas.

En donde el símbolo # se conoce como "carta suplementaria" e indica los argumentos que van incluidos dentro de la instrucción.

2.2.4. SPSS

SPSS es un software estadístico privativo que brinda al usuario una amplia gama de capacidades estadísticas y analíticas con el fin de facilitar el manejo de los datos y la interpretación de la información, para de esta forma enriquecer el proceso de toma de decisiones en el mundo de los negocios.
Es desarrollado por la compañía Estadounidense líder en tecnología IBM (International Business Machines), la cual lo define como un software que "pone el poder del análisis estadístico avanzado en sus manos".

Dentro de las ventajas principales que ofrece SPSS se destaca por un lado su estructura de funcionamiento: trabaja en todos los sistemas operativos, tipos de archivos y datos y lenguajes externos de programación; y por otro lado su capacidad de amoldamiento: proporciona funcionalidades e interfaces personalizadas de acuerdo a las responsabilidades y diferentes niveles de habilidad de los usuarios: empresarios, analistas y estadísticos.

Es importante mencionar que SPSS Statistics trabaja por defecto con archivos de extensión *.sav,* sin embargo tiene la ventaja de reconocer y producir archivos que puedan ser leídos por otros programas tales como Excel, R-Project, SAS y Stata (a los cuales corresponde respectivamente las extensiones: *.xls* o *.xlsx, .csv, .sd2 y .dta)*.

La interfaz del software está divida en dos grandes partes: la cinta de opciones, ubicada en la parte superior, contiene todas las funciones que ofrece SPSS, las cuales aparecen en menús desplegables y algunos íconos que representan a las más importantes. Y el área central de trabajo la cual es similar a una hoja de cálculo de Excel y se encuentra dividida en *Vista de Datos* y *Vista de Variables*; la primera está organizada en filas y en columnas, en las cuales aparecen respectivamente las Variables y las Observaciones, y es la vista que nos permite ingresar, modificar y eliminar los valores; la segunda nos permite definir una serie específica de parámetros y características sobre las variables.

Para comenzar a utilizar el software, el usuario puede escribir los datos directamente en el área de trabajo o puede importar alguna base de datos en la cual haya trabajado previamente, para lo cual debe dirigirse a ***Archivo > Abrir > Datos*** y seleccionar la

ubicación del archivo (allí podrá seleccionar el formato del archivo con el cual desea trabajar).

> *Nota:* Para efectos de demostración de las herramientas del software, se va a trabajar con una base de datos denominada Datos de Longley, la cual es reconocida en el área de econometría por mostrar una alta multicolinealidad (supuesto de los modelos de regresión que será expuesto más adelante).

El usuario notará que luego de abrir la base de datos, tanto la vista de datos como la de variables se llenará automáticamente e igualmente una ventana de resultados aparecerá. En esta última se mostrarán todas las instrucciones y operaciones que el usuario realice sobre el archivo.

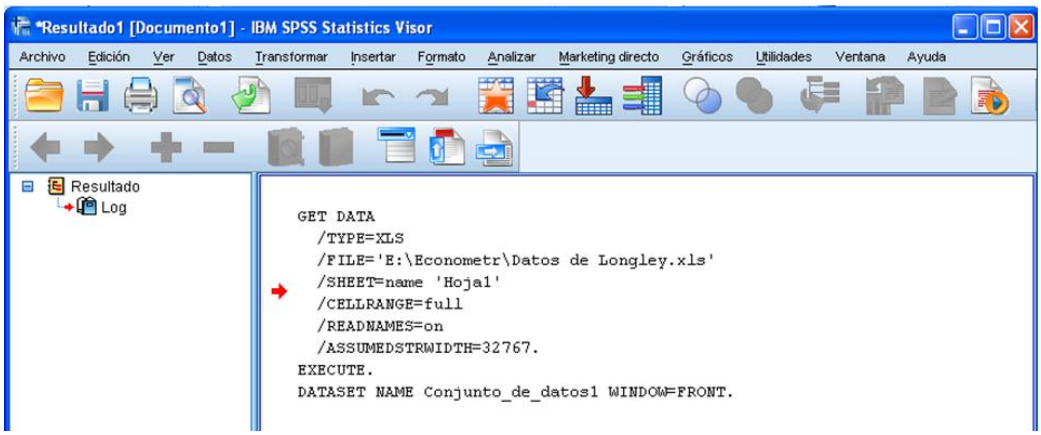


Ilustración 18. Ventana de resultados-SPSS

Dado que SPSS es un software estadístico, las herramientas de análisis que ofrece son muy completas; especialmente la estadística descriptiva la cual se puede realizar sin necesidad de un código, sino simplemente siguiendo unos sencillos pasos:

1. En la pestaña *Analizar* que se encuentra ubicada en la barra de opciones, seleccionar ***Estadísticos Descriptivos > Descriptivos.***

2. Se abre una ventana en la cual se deberá especificar las variables y el tipo de análisis que se desea realizar (por tipo de análisis nos referimos a los estadísticos descriptivos que desee incluir). Para escoger las variables puede efectuar dos operaciones: seleccionarlas y arrastrarlas hasta el cuadro "variables" ó haciendo clic en la flecha que se encuentra en la mitad de los dos cuadros. (Para remover una variable del análisis podrá hacer lo mismo).

3. Para elegir los estadísticos descriptivos que desea incluir en el análisis debe dirigirse al botón "**Opciones**".

Ilustración 19. Elección de estadísticos descriptivos-SPSS

4. Finalmente deberá dar ***Aceptar*** e inmediatamente el proceso se generará en la ventana de resultados. De esta forma podrá tener toda la información necesaria sobre el comportamiento de los datos en una sola tabla, enseñada en la ilustración 16.

Descriptivos

[Conjunto_de_datos1]

Estadísticos descriptivos

	N	Rango	Mínimo	Máximo	Media	Desv. típ.	Varianza	Asimetría		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Error típico	Estadístico	Error típico
employed	16	10380	60171	70551	65317,00	3511,968	12333921,73	-,104	,564	-1,399	1,091
GNPdeflator	16	34	83	117	101,68	10,792	116,458	-,162	,564	-1,151	1,091
GNP	16	320605	234289	554894	387698,44	99394,938	9,879E9	,028	,564	-1,072	1,091
unemployed	16	293,6	187,0	480,6	319,331	93,4464	8732,234	,175	,564	-,998	1,091
armedforces	16	214	146	359	260,67	69,592	4843,041	-,448	,564	-,835	1,091
population	16	22473	107608	130081	117424,00	6956,102	48387348,93	,319	,564	-,950	1,091
N válido (según lista)	16										

Ilustración 20. Output Descriptivos-SPSS

A partir de lo expuesto anteriormente es posible observar que todos los software analizados ofrecen una amplia gama de herramientas estadísticas que permiten analizar con más detalle el comportamiento de los datos, por lo que inicialmente no es posible concluir acerca de la mayor o menor utilidad o ventajas y desventajas de algún software en especial. Sin embargo, este proceso de comparación entre las diferentes herramientas

proporcionadas, se reduce directamente a la exploración de las capacidades de los software frente a nuestra temática central: el método de regresión lineal.

Es por esto que surge la necesidad de hacer una breve exposición acerca del método, los supuestos que lo componen y realizar luego una revisión detallada de los software (similar a la previamente observada) para finalmente hacer una aplicación práctica que ilustre todas las conclusiones a las cuales se ha de llegar.

2.3. MÉTODO DE REGRESIÓN LINEAL

Como ya se ha expresado en secciones anteriores, la regresión lineal es un método de análisis estadístico que no se aplica únicamente en las Ciencias Económicas, al contrario, es una técnica ampliamente utilizada en los diferentes campos y disciplinas de las Ciencias Sociales y Naturales debido a las ventajas que ofrece en cuanto a la realización de análisis estructurales, predicciones de valores futuros y evaluación de políticas, entre otras.

En términos más generales, un modelo de regresión se emplea para obtener una descripción y evaluación de la posible relación existente entre una variable llamada endógena (Y) y una o más variables llamadas exógenas (X); conocidas igualmente como variable dependiente e independiente respectivamente. Si tiene una sola variable exógena se denomina regresión simple y si tiene dos o más exógenas, regresión múltiple[13] (Ilustración 21).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + U_i$$

Ilustración 21. Regresión múltiple

Adicionalmente, es importante resaltar que el método de regresión lineal, como su nombre lo indica, hace referencia a la linealidad de los parámetros β[14] más no necesariamente de las variables, las cuales pueden estar en cualquier forma lineal.

De esta forma, el objetivo de un modelo de regresión es estimar la Función de Regresión Muestral que sea lo más parecida posible a la Función de Regresión Poblacional a partir de una muestra de datos, lo cual se logra por medio de la estimación de los parámetros β tal que se minimice la suma de los residuos al cuadrado. Estos parámetros estimados se

[13] Un modelo de regresión en su totalidad se considera aleatorio gracias al término de error (U).

[14] Existen casos particulares donde la función no es lineal pero se puede linealizar por medio del logaritmo.

conocen como los estimadores de Mínimos Cuadrados Ordinarios o MCO (Ordinary Less Squared), los cuales cumplen con las propiedades de ser lineales, son una combinación lineal de una variable aleatoria; insesgados, el valor esperado del estimador es igual al verdadero parámetro poblacional y de varianza mínima.

En las diferentes áreas en donde se utiliza el método de regresión lineal, a los investigadores les interesa saber el tipo de relación que pueden encontrar entre diversas variables, razón por la cual vamos a centrar nuestra atención en el modelo de regresión múltiple, el cual se trabaja de forma matricial y está expresado por la siguiente ecuación,

$$Y = X\hat{\beta} + \hat{U}$$

En donde Y representa el vector de la variable endógena (tamaño n*1), X la matriz de variables exógenas (tamaño n*k), $\hat{\beta}$ el vector de los parámetros beta (tamaño k*1) y $\hat{U}$ el vector de los errores (tamaño n*1).

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & X_{12} & X_{13} & \dots & X_{1k} \\ 1 & X_{22} & X_{23} & \dots & X_{2k} \\ 1 & X_{32} & X_{33} & \dots & X_{3k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n2} & X_{n3} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \hat{U}_3 \\ \vdots \\ \hat{U}_4 \end{bmatrix}$$

La columna de 1 representa los términos independientes

Ilustración 22. Regresión múltiple en términos matriciales

De acuerdo a la condición establecida previamente, la ecuación que permite encontrar los parámetros $\hat{\beta}$ tal que se minimice la suma de residuos al cuadrado está definida por la expresión,

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Ilustración 23. Estimador de Mínimos Cuadrados Ordinarios

2.3.1. SUPUESTOS DEL MODELO DE REGRESIÓN

Para construir, estimar y poder aplicar correctamente un modelo de regresión lineal es necesario que cumpla con una serie de supuestos, los cuales aparecen listados a continuación.

1. El modelo debe ser lineal en los parámetros.

2. El valor esperado del vector de residuos es un vector nulo, es decir que la media de los residuos es igual a cero.

$$E(U) = 0$$

3. La varianza de los residuos debe ser constante a lo largo de la muestra. Este se conoce como el supuesto de HOMOSCEDASTICIDAD.

$$\text{Var}\,(U_i) = \sigma_u^2 \qquad \text{para todo i}$$

4. Debe existir independencia entre los residuos de un periodo con los de otro u otro periodos; esto equivale a decir que los residuos sean independientes o que su covarianza sea igual a cero. Este se conoce como el supuesto de NO AUTOCORRELACIÓN. [15]

$$\text{Cov}\,(U_i U_j) = 0 \qquad \text{para todo } i \neq j$$

5. Los residuos deben seguir una distribución normal, es decir deben tener media cero y varianza σ^2.

$$U \sim N(0, \sigma^2)$$

6. Debe existir independencia lineal entre las variables exógenas del modelo, es decir el rango de la matriz X es completo. Este se conoce como el supuesto de NO MULTICOLINEALIDAD.

$$r\,(X) = K \qquad \text{donde } n > k$$

7. Se supone que los $\hat{\beta}$, o los coeficientes de regresión estimados permanecen constantes a lo largo de la muestra, es decir NO HAY CAMBIO ESTRUCTURAL y hay estabilidad de los parámetros.

8. Debe existir independencia entre las variables exógenas y los residuos del modelo. En otros términos, la covarianza entre los residuos y las exógenas debe ser cero.

Al momento de construir un modelo de regresión utilizando una muestra de datos aleatorios, no se tiene total certeza de que éste cumple todos los supuestos y por lo tanto no

[15] Los supuestos de homoscedasticidad y autocorrelación se resumen en que la matriz de Var-Cov debe ser escalar: $E(UU') = \sigma_u^2 I$

puede haber confiabilidad en que la aplicación del modelo va a producir resultados lógicos y coherentes con lo observado o con la teoría económica. Por esta razón es necesario realizar una cierta cantidad de pruebas sobre el modelo que permiten verificar dichos supuestos; y en dado caso que no se cumplieran brindan varías opciones para corregir los errores que se presenten con el fin de que el modelo sea lo más exacto posible.

2.4. REGRESIÓN LINEAL Y VERIFICACIÓN DE SUPUESTOS

En primera instancia se van a considerar los supuestos que tienen que ver con la estructura del modelo de regresión, entre los cuales se encuentran las hipótesis de Muestras pequeñas, Cambio Estructural, Especificación errónea y Multicolinealidad. El hecho de que el modelo no cumpla con cada uno de estos supuestos produce consecuencias negativas sobre la exactitud de la estimación, haciendo que el modelo se aleje más de la realidad, por ende es necesario evitar y corregir en lo posible su violación.

Por otro lado, al igual que en el apartado anterior, se pretende hacer una revisión de las diferentes opciones y facilidades que cada software ofrece para cumplir con el análisis pertinente. Igualmente para poder realizar esto se deberá hacer una exposición de la forma en la cual cada software permite realizar una regresión lineal.

2.4.1. DESARROLLO DEL MÉTODO DE REGRESIÓN EN STATA, R-PROJECT, RATS Y SPSS

Nota: *Para efectos de simplificación y comparación de resultados, en todos los software se va a trabajar con la misma base de datos utilizada en la sección anteriormente dedicada a WinRATS.*

2.4.1.1. Stata 11.0

Luego de haber cargado los datos en Stata se puede proceder directamente a realizar la regresión.
La instrucción a ejecutar es ***reg*** *o* ***regres*** y debe ir seguida por la variable dependiente y las variables independientes en orden respectivo. Para la base de datos que estamos utilizando se obtuvieron los siguientes resultados:

```
. reg inverex pibr e inf ir
```

Source	SS	df	MS
Model	8.3392e+13	4	2.0848e+13
Residual	7.6473e+12	51	1.4995e+11
Total	9.1039e+13	55	1.6553e+12

Number of obs	= 56
F(4, 51)	= 139.04
Prob > F	= 0.0000
R-squared	= 0.9160
Adj R-squared	= 0.9094
Root MSE	= 3.9e+05

inverex	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pibr	.7472896	.0335744	22.26	0.000	.6798862	.8146931
e	12033.07	7936.616	1.52	0.136	-3900.346	27966.49
inf	198847.4	17585.36	11.31	0.000	163543.3	234151.4
ir	-25237.95	11147.82	-2.26	0.028	-47618.14	-2857.764
_cons	-1.41e+07	1469431	-9.62	0.000	-1.71e+07	-1.12e+07

Ilustración 24. Regresión-Stata

Los resultados muestran los coeficientes de los diferentes parámetros, su error estándar, el estadístico t, el valor P (probabilidad), los intervalos de confianza y otros valores importantes como el R-cuadrado y la prueba F para significancia global de los parámetros. Adicionalmente l atabla que se encuentra en l pate superior izquierda se conoce como ANOVA y contiene la suma de residuos al cuadrado (SS), los grados de libertad (df) y el promedio de la suma de residuos al cuadrado (MS).

El proceso se puede realizar igualmente a través de las pestañas del software. En la pestaña ***Statistics>Linear Models and related>Linear Regression***. Aparecerá una ventana en la cual deberá especificar los argumentos para la regresión: variable dependiente y variables independiente.

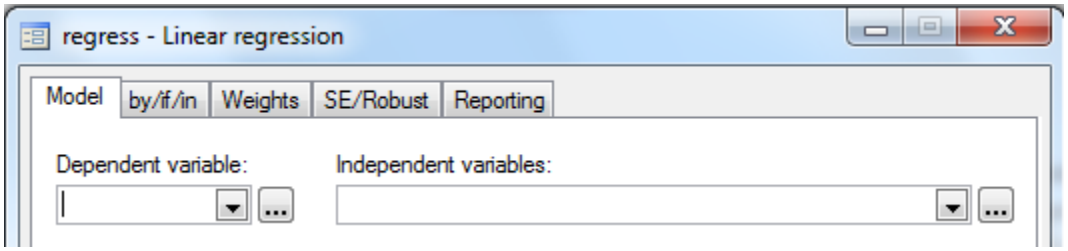


Ilustración 25. Regresión alterna-Stata

2.4.1.2. R-Project

Para correr una regresión en R, el comando utilizado es **lm()**, el cual sirve en términos generales para ajustar modelos lineales (lineal models) lo que significa que no es útil únicamente para hacer regresión lineal, sino también análisis de varianza y covarianza. El

comando lleva dentro del paréntesis la variable endógena acompañada del símbolo ~ , seguido de las variables exógenas separadas por el signo + y por último, separado por una coma, el nombre que se le dio al archivo sobre el cual se trabaja. En el ejemplo a seguir, el comando se empleó según la ilustración 26:

```
> multiple <- lm(INVEREX~PIBR+E+INF+IR, Base1)
```

Ilustración 26. Regresión-R

Cuando se llama el objeto, el resultado inmediato que presenta el software son los coeficientes que acompañan a cada variable de la regresión. Sin embargo, como se dijo anteriormente, a través del comando **summary()** se puede obtener información más detallada.

```
> summary(multiple)

Call:
lm(formula = INVEREX ~ PIBR + E + INF + IR, data = Base1)

Residuals:
    Min      1Q  Median      3Q     Max
-932883 -262370   -5898  268322  891382

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.414e+07  1.469e+06  -9.624 4.69e-13 ***
PIBR         7.473e-01  3.357e-02  22.258  < 2e-16 ***
E            1.203e+04  7.937e+03   1.516   0.1357
INF          1.988e+05  1.759e+04  11.308 1.66e-15 ***
IR          -2.524e+04  1.115e+04  -2.264   0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 387200 on 51 degrees of freedom
Multiple R-squared: 0.916,      Adjusted R-squared: 0.9094
F-statistic:   139 on 4 and 51 DF,  p-value: < 2.2e-16
```

Ilustración 27. Resumen de regresión-R

Con esto se especifican por un lado algunos datos estadísticos sobre los residuos del modelo estimado; por otro lado, además de los coeficientes anteriormente obtenidos, muestra el error estándar, el valor t de la distribución y el valor p o la probabilidad que resulta muy útil a la hora de hacer pruebas de hipótesis y de significancia individual y global (Es importante resaltar que R incluye las convenciones para evaluar la significancia respecto a ciertos niveles). Y en último lugar enseña el R-cuadrado (o coeficiente de determinación), el cual mide en qué porcentaje los variables exógenas del modelo explican

la variabilidad de la variable endógena, y el R-cuadrado ajustado el cual se usa de forma específica para comparar modelos alternativos.

2.4.1.3. WinRATS 7.2.

El comando a utilizar en este software está especificado de la siguiente manera:

```
linreg y / error
# constant x1 x2 x3 x4
```

El comando propiamente es **linreg** y va acompañado al lado derecho por la variable dependiente y otro término que permite visualizar información sobre los errores del modelo. En la parte inferior, se encuentran las variables explicativas del modelo al lado del símbolo # el cual representa los argumentos que se incluyen en la acción a desarrollar; esta última parte es conocida como una carta suplementaria.

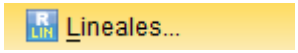
```
Linear Regression - Estimation by Least Squares
Dependent Variable Y
Quarterly Data From 1994:01 To 2007:04
Usable Observations     56      Degrees of Freedom    51
Centered R**2     0.916000      R Bar **2   0.909411
Uncentered R**2   0.992129      T x R**2      55.559
Mean of Dependent Variable      3965267.5577
Std Error of Dependent Variable 1286568.8408
Standard Error of Estimate       387230.5839
Sum of Squared Residuals       7.64732e+012
Regression F(4,51)                139.0352
Significance Level of F         0.00000000
Log Likelihood                  -797.38126
Durbin-Watson Statistic           0.656175

  Variable      Coeff           Std Error       T-Stat      Signif
************************************************************************
1. Constant   -14141332.9   1469431.14   -9.62368  0.00000000
2. X1             0.74729      0.033574    22.2577  0.00000000
3. X2         12033.0725   7936.61624    1.51615  0.13565657
4. X3         198847.355   17585.3591   11.30755  0.00000000
5. X4        -25237.9498   11147.8242   -2.26394  0.02786106
```

Ilustración 28. Regresión-RATS

Los resultados expuestos por RATS con muy similares a los expuestos por los anteriores software expuestos, la diferencia más significativa es la organización de la información: en la parte superior se encuentra una lista compuesta sobre información y estadísticos básicos de la regresión y después una tabla dedicada específicamente a mostrar los valores de los estimadores y su significancia.

2.4.1.4. SPSS

El desarrollo de la regresión en SPSS se hace sin utilización de códigos, simplemente mediante selecciones de pestañas y botones. En la pestaña *Analizar* se encuentra la opción *Regresión* y dentro de ésta el ícono 

. Luego de seleccionar el ícono, emerge una ventana solicitando los argumentos de la regresión.

Ilustración 29. Regresión-SPSS

En la parte izquierda se encuentra la lista de variables y en el medio los campos para agregar dichas variables de acuerdo a su función: dependientes, independientes. SPSS ofrece adicionalmente la utilización de tres campos ubicados en la parte inferior de la ventana, los cuales permiten en su respectivo orden: elegir una variable de selección para limitar el análisis a un subconjunto de casos, seleccionar una variable de identificación de

casos para identificar los puntos en los diagramas y seleccione una variable numérica de Ponderación MCP para el análisis de mínimos cuadrados ponderados[16].

En la parte derecha de la ventana, aparecen unos botones que nos permiten agregarle detalles y estadísticos importantes a nuestra regresión. En Estadísticos podrá agregar elementos como intervalos de confianza, matriz de covarianzas y otros importantes para el análisis de los parámetros; por medio de las opciones Gráficos y Opciones es posible generar variedad de gráficos para cada variable y configurar pequeños valores de los estadísticos.

La regresión observada en la ventana de resultados se encuentra organizada por tablas que representan el resumen del modelo, la ANOVA (Análisis de Varianzas), los coeficientes y la correlación entre ellos. En este caso nos interesa únicamente observar la composición del resumen del modelo y los coeficientes.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	-14141332,88	1469431,144		-9,624	,000	-17091338,99	-11191326,76
	PIBR	,747	,034	1,334	22,258	,000	,680	,815
	E	12033,073	7936,616	,119	1,516	,136	-3900,349	27966,494
	INF	198847,355	17585,359	1,053	11,308	,000	163543,274	234151,437
	IR	-25237,950	11147,824	-,124	-2,264	,028	-47618,141	-2857,759

a. Variable dependiente: INVEREX

Ilustración 30. Output regresión-SPSS

En la Ilustración 30 encontramos más detalladamente los resultados de la regresión: los coeficientes de los estimadores y los estadísticos que permiten hacer conclusiones sobre su significancia individual.

[16] Información tomada del tutorial de SPSS.
http://127.0.0.1:56593/help/index.jsp?topic=/com.ibm.spss.statistics.tut/introtut2.htm

2.5. SUPUESTOS QUE TIENEN QUE VER CON LA ESTRUCTURA DEL MODELO

2.5.1. HIPÓTESIS DE MUESTRAS PEQUEÑAS

El hecho de que el modelo de regresión lineal se efectué sobre muestras de tamaño pequeño no afecta las propiedades de los estimadores que de este se derivan. Las consecuencias de esto se pueden ver expresadas en una secuencia: todo comienza por que la existencia de muestras pequeñas hace que la Varianza de las variables exógenas aumente en gran proporción[17]; por otro lado se evidencia que la varianza de los residuos y de los parámetros β también aumenta, lo que lleva a aumentar de igual forma su desviación estándar.

Esto finalmente tiene dos efectos negativos sobre el modelo: por una parte hace que los intervalos de confianza[18] se vuelvan mucho más amplios perdiendo confiabilidad, y por la otra parte hace que las pruebas de significancia individual sobre los parámetros indiquen que las variables que los acompañan no son significativas para el modelo, cuando en realidad sí lo son.

En realidad no existen pruebas para determinar si un modelo proviene de muestra grande o pequeña, no hay un número exacto a partir del cual se pueda hacer la distinción; sin embargo con los efectos que las muestras pequeñas tienen sobre los resultados del modelo, es posible a partir de la obtención de resultados no lógicos, sospechar que el modelo no cuenta con una muestra lo suficientemente grande para ser exacto y fiel en sus conclusiones.

La solución más conocida para evitar este tipo de problemas es tratar de volver el modelo parsimonioso, es decir explicar la variable endógena con el menor número de variables exógenas posibles; efecto que contrarresta lo demás.

[17] Debido a la forma en como esta es calculada: $Var(X) = \frac{\sum(X_i - \bar{X})^2}{n-1}$

[18] Intervalo calculado en el cual se encuentra el valor del parámetro.

2.5.2. HIPÓTESIS DE CAMBIO ESTRUCTURAL

Aunque el significado específico de Cambio estructural depende en gran manera del entorno en el cual se esté trabajando, para efectos del modelo de regresión lineal puede ser entendido como la existencia de un cambio significativo en la estructura del modelo durante el periodo de observación, el cual puede ser producido por diversos choques externos en alguna/s variable.

La existencia de un cambio estructural puede producir, además de causar inestabilidad en los parámetros, tres efectos negativos:

1. Si se presenta una situación de cambio estructural y no es tenida en cuenta para la estimación del modelo, los estimadores obtenidos van a ser sesgados respecto al comportamiento de cada una de las estructuras diferentes del modelo, es decir sesgados respecto a los verdaderos estimadores que se obtendrían si se tuviera en cuenta el cambio estructural.

2. La suma de residuos al cuadrado va a ser mucho mayor (a la que debería ser) en el modelo en donde no se tenga en cuenta el cambio estructural. Esto va a generar un efecto muy similar al de la existencia de muestras pequeñas: la varianza de los residuos y de los parámetros será muy grandes, haciendo que los intervalos de confianza y las pruebas de significancia pierdan confiabilidad y se presenten resultados incoherentes con la teoría o la realidad.

3. Si no se tiene en cuenta la presencia de cambio estructural, el modelo puede aparentar problemas de Heteroscedasticidad o de Autocorrelación.

La pregunta que surge entonces es cómo detectar con exactitud el momento en el cual se produjo el cambio estructural. Para este efecto, existen tres técnicas ampliamente conocidas: el método gráfico, el Test de Chow y el Test de Cusum. El método gráfico es considerado como una medida muy subjetiva ya que difícilmente puede proporcionar información al investigador acerca del momento exacto en el cual se produce el cambio, sin embargo puede resultar útil como una primera aproximación.

Test de Chow

El Test de Chow es una reconocida prueba utilizada para detectar la existencia de cambio estructural, para lo cual sugiere una serie de pasos:

- Primero estimar un modelo de regresión como si no hubiese cambio estructural y calcularle la Suma de Residuos al Cuadrado (SRC).
- Hacer una regresión para la primera sub-muestra, es decir la muestra antes de que ocurriera el cambio estructural, e igualmente calcularle la Suma de Residuos al Cuadrado (SRC_1).
- Hacer otra regresión para la muestra después de ocurrido el cambio estructural (incluyendo el momento en el que ocurre) y calcular la Suma de Residuos al Cuadrado (SRC_2).[19]
- Por último aplicar una prueba de hipótesis y contrastarla con una prueba F[20] como se ilustra a continuación:

$H_0 : \beta_0 = \beta_1 = \beta_n$ → No hay cambio estructural

$H_1 : \beta_0 \neq \beta_1 \neq \beta_n$ → Hay presencia de cambio estructural

Prueba F: $$F = \frac{\frac{SRC - (SRC_1 + SRC_2)}{k}}{\frac{(SRC_1 + SRC_2)}{n-2k}} \sim F_\alpha\,(k, n-2k)$$

Donde cada uno de los argumentos de la parte derecha son los grados de libertad y k es el número de parámetros del modelo.

Si el estadístico calculado es mayor al estadístico de la tabla (parte izquierda), se rechaza la hipótesis nula y se dice que hay presencia de cambio estructural; de lo contrario no hay. En el caso de los software, es de mayor utilidad sacar conclusiones sobre el valor-p: si éste es menor al valor α (regularmente asumido como 0.05) se rechaza la hipótesis nula, si el valor-p es mayor, no se rechaza.

De esta forma, el investigador podrá tener más certeza sobre la existencia o no de cambio estructural. Sin embargo, el Test de Chow tiene algunos inconvenientes y limitantes que necesitan ser corregido para su correcta aplicación. En primer lugar, se necesita saber a priori el posible punto del cambio estructural, lo cual se puede solucionar mediante el conocimiento del comportamiento histórico o gráfico de los datos. Adicionalmente, si el cambio

[19] Cada submuestra debe tener un tamaño mínimo: el número de observaciones debe exceder al número de parámetros.

[20] Prueba en donde el estadístico utilizado sigue una distribución F.

estructural se acerca a uno de los extremos de la muestra, la prueba pierde potencia y el estadístico calculado debe ser corregido,

$$F = \frac{\frac{SRC - SRC_{Muestra\ Grande}}{n_{MuestraPequeña}}}{\frac{SRC_{Muestra\ Grande}}{n_{MuestraGrande} - k}} \sim F_{\alpha}\ (n_{MuestraPequeña}, n_{MuestraGrande} - k)$$

Test de CUSUM

El Test de Cusum se utiliza igualmente para evaluar la presencia de cambio estructural, permitiendo saber con mayor exactitud el momento en el cual puede existir un cambio estructural (solucionando una limitación de Chow).
El Test utiliza una hipótesis nula similar a la del Test de chow, en la cual a partir de los residuos, los parámetros se consideran estables y una hipótesis alternativa la cual indica que los parámetros son constantes hasta cierto momento t*.

A partir de eso, es posible analizar el comportamiento gráfico de esa suma de residuos generados y calcular o establecer unos límites de confianza (representados gráficamente por bandas) dentro de los cuales se va a encontrar la curva. Se dice que si el gráfico sobrepasa en algún punto esos límites de confianza, se está sugiriendo que en ese punto hay inestabilidad de los parámetros y por lo tanto presencia de cambio estructural.

Como alternativa adicional, puede utilizarse la prueba CUSUM cuadrado, la cual realiza el mismo proceso pero partiendo de la suma de residuos al cuadrado y evidenciando posibles desviaciones respecto a su valor medio.

2.5.2.1. APLICACIÓN EN SOFTWARE

2.5.2.1.1. Stata 11.0

El procedimiento para probar la estabilidad de los parámetros en Stata consiste en desarrollar por separado todos los pasos del test de Chow explicados anteriormente. En primer lugar se estima una regresión normal para el modelo observando el valor de la Suma de Residuos al Cuadrado, luego es necesario dividir las observaciones en dos sub-muestras, una antes del cambio estructural y otra después de ocurrido. Para separar de

forma fácil y rápida la muestra, se debe hacer la inclusión de una variable año que contenga los períodos. Como la base de datos tiene observaciones trimestrales, es necesario repetir el mismo año durante cuatro observaciones, es decir 1994, 1994, 1994, 1994, 1995,1995,1995,1995,… y así sucesivamente; en este caso esa variable fue denominada YEAR. Para poder efectuar las regresiones de forma independiente para cada sub-muestra es necesario adicionar al comando **reg** el comando **if** el cual está diseñado para efectuar funciones de forma restringida sobre ciertas observaciones. De esta forma se harán dos regresiones adicionales incluyendo el comando para separar el período anterior y posterior al cambio[21].

```
. reg  inverex pibr e inf ir if year<2002
. reg  inverex pibr e inf ir if year>=2002
```

Ilustración 31. Regresión sub-muestras-Stata

Para cada sub-muestra se obtienen los resultados de una regresión, de los cuales para efectos del cálculo del estadístico F se necesita la Suma de Residuos al Cuadrado. En el ejemplo se obtuvo lo siguiente:

SRC: 7.6473e+12 SRC_1: 3.1885e+12 SRC_2: 9.6377e+11

Con esa información, el estadístico F calculado es 7.743. Lo siguiente que debe hacerse para revisar el resultado del test, es recurrir a la tabla del estadístico F a revisar el valor crítico que toma al nivel de confianza y grados de libertad determinados; esto para poder comparar el valor calculado y concluir. El valor crítico del estadístico F en la tabla corresponde a $F_{0.05}\,(5{,}46) = 0.22$, de ésta forma se concluye que en el primer trimestre del 2002 se presenta cambio estructural.

Por otro lado, como se vio anteriormente, otra alternativa para evaluar la inestabilidad de los parámetros es el test de CUSUM. Stata cuenta con un paquete llamado **cusum6** que ejecuta la prueba de forma automática; para descargarlo se utiliza el comando **ssc install** seguido del nombre del paquete, lo que produce el siguiente resultado:

```
. ssc install cusum6
checking cusum6 consistency and verifying not already installed...
installing into c:\ado\plus\...
installation complete.
```

Ilustración 32. Instalación paquetes-Stata

[21] Para efectos del ejemplo se desea probar si hubo cambio estructural en el primer período del 2002 correspondiente a la observación 33.

Luego de instalado el paquete, podemos hacer uso de la función; ésta, requiere la existencia de una variable del tipo serie de tiempo, la cual puede ser establecida por medio del comando **tsset**.

```
. tsset  year
        time variable:  year, . to .
                delta:  1 unit
```

Ilustración 33. Variable tipo serie de tiempo-Stata

Nota: Para caracterizar así la variable tiempo, es necesario que no se repitan datos en la muestra por lo que la base de datos empleada para aplicar el test de Chow debe ser modificada; esto se puede hacer con el comando **replace** o cambiando los datos manualmente por medio del Editor de Datos (Window>Data Editor).

A continuación se puede proceder a la aplicación del test de CUSUM, utilizando para esto el comando anteriormente mencionado y acompañado de las variables del modelo. Inmediatamente después de ejecutada la instrucción, se genera en una nueva ventana la gráfica de CUSUM y al seleccionar la opción more ubicada debajo del comando se genera la gráfica CUSUM Cuadrado.

```
. cusum6 inverex pibr e inf ir year
--more--
```

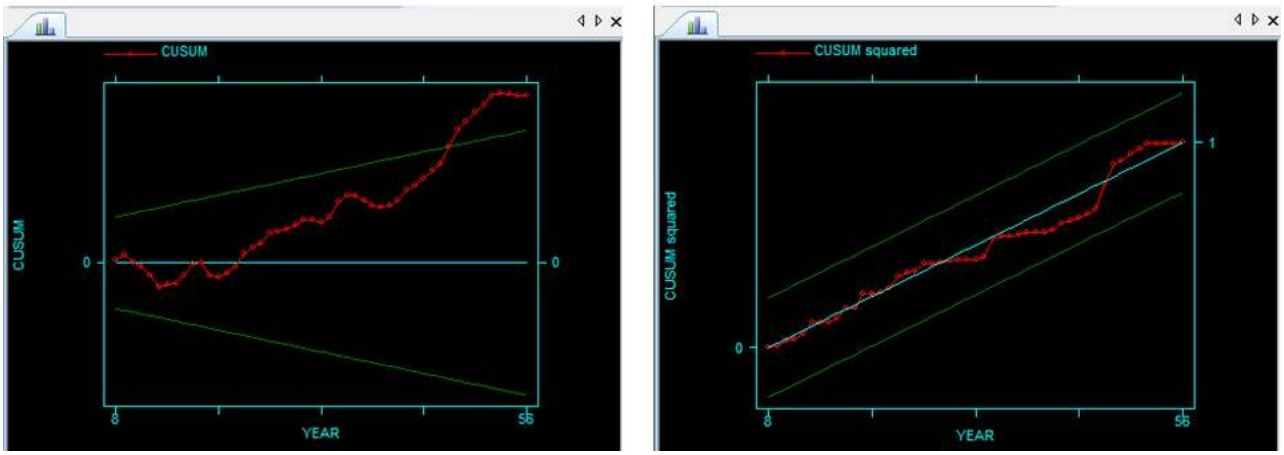


Ilustración 34. CUSUM-Stata

2.5.2.1.2. R-Project

R-project ofrece a los usuarios una amplia serie de alternativas para evaluar la presencia de cambio estructural en los modelos de regresión lineal. Para esto, se trabaja por medio del paquete **strucchange**[22], el cual está diseñado específicamente para ese fin.

Luego de descargado el paquete, lo primero que se debe hacer es llamarlo mediante el comando

```
> library(strucchange)
```

Esto nos permitirá hacer uso de todas sus funciones. Dentro de éstas, la función utilizada para detectar el cambio estructural es **sctest,** la que a su vez ofrece dos pruebas a aplicar: Chow y Nyblom-Hansen. Los principales argumentos que debe contener la función son[23]:

- Formula: una fórmula que decriba el modelo sobre el cual va a ser aplicada la prueba.
- Type: Una cadena de caracteres que especifica la prueba que se va a realizar: "Chow" o "Nyblom-Hansen"
- Point: Parámetro de la prueba de Chow para ubicar el punto potencial del cambio estructural.

Antes de proceder con la prueba, es necesario digitar el comando **attach()**, incluyendo dentro del paréntesis el nombre dado al conjunto de datos sobre el cual se trabaja; esto con el fin de que R reconozca individualmente cada una de las variables del modelo y así evitar problemas en el momento de especificar la función. Para el ejemplo, la función se aplicó de la siguiente manera:

```
> sctest(INVEREX ~ PIBR + E + INF + IR, data = rh,type = "Chow",point=10)
```

Ilustración 35. Test de Chow-R

El primer argumento representa la fórmula que identifica al modelo, *data* es un argumento adicional el cual está especificando el conjunto de datos del cual se obtienen las variables (si no se incluye, las variables se toman por defecto del ambiente desde el cual se aplica la función), el tercer argumento corresponde a la identificación puntual de la prueba

[22] El paquete podrá descargarlo de http://cran.R-project.org/

[23] Podrá consultar otros argumentos a incluir en la documentación del software.

a realizar y el último equivale al punto en el cual se cree existe cambio estructural (en este caso fue tomado aleatoriamente). Los resultados obtenidos fueron:

```
        Chow test

data:  INVEREX ~ PIBR + E + INF + IR
F = 16.5913, p-value = 2.422e-09
```

Dentro del resultado expuesto por el software, se identifican el valor calculado del estadístico F y el valor-p, sobre lo cual se puede concluir que en el momento 10 hay presencia de cambio estructural.

Por su lado, para realizar la prueba de CUSUM R-project trabaja a partir de pruebas de fluctuación generalizada, las cuales tienen como objetivo derivar del modelo de regresión procesos empíricos que capturen las fluctuaciones ya sea en las estimaciones o en los residuales; de esta forma, cuando se presenten grandes fluctuaciones se rechazará la hipótesis de estabilidad de los parámetros y por lo tanto habrá presencia de cambio estructural.

En R, este proceso se lleva a cabo mediante la función **efp** (Empirical fluctuation processes), la cual contempla pruebas como CUSUM y MOSUM que están basadas en los residuales recursivos o los obtenidos por el método de MCO. Al igual que las anteriores funciones utilizadas, **efp** incluye dos argumentos de gran importancia: el primero se refiere a la fórmula que representa al modelo de regresión y el segundo al tipo de prueba que se va a aplicar; en este caso CUSUM basada en los residuales obtenidos por mínimos cuadrados (OLS-CUSUM).

```
> cus <- efp(INVEREX~PIBR+E+INF+IR, data=rh, type="OLS-CUSUM")
```

Esta función arroja un objeto de clase 'efp' que contiene el proceso empírico de fluctuación. Debido a que el análisis de la prueba de CUSUM se hace a partir del comportamiento gráficos de esos residuos, ese necesario implementar dos fórmulas más. La primera está relacionada con el establecimiento de los límites de confianza dentro de los cuales se va a evaluar la curva; para fijar esos límites, se utiliza la función **boundary()**, acompañada dentro del paréntesis por el nombre del objeto que representa la función 'efp' y un nivel de significancia α, por defecto 0.05.

```
> bound.cus <- boundary(cus, alpha=0.05)
```

Por último es necesario gráficar el comportamiento de la función efp, la cual ya incluye los límites, por medio de la función **plot()**. El resultado obtenido sugiere un posible intervalo de tiempo donde puede presentarse cambio estructural, el cual se identifica con la curva que se encuentra por fuera del límite superior.

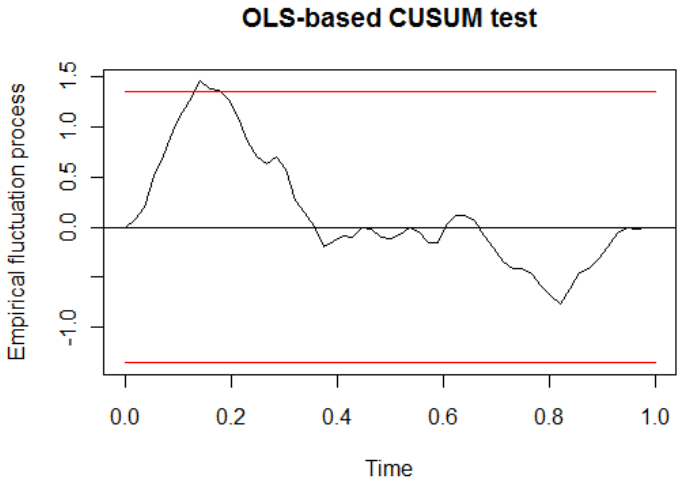


Ilustración 36. Resultado CUSUM-R

2.5.2.1.3. WinRATS 7.2

RATS, al igual que los anteriores software ofrece la posibilidad de aplicar el test de Chow y el test de CUSUM para evaluar la presencia de cambio estructural. Para el primer test, el proceso a realizar es muy similar al expuesto en la sección de Stata: primero se hace la regresión para toda la muestra, luego, se hace por separado para la primera sub-muestra y para la segunda, y finalmente se calcula el estadístico F.

Para calcular las regresiones para cada sub-muestra es necesario adicionar al comando de regresión los años durante los cuales se está llevando a cabo el periodo de análisis, como se ilustra

```
linreg(noprint) y 1994:01 2001:04
#constant x1 x2 x3 x4
```

El comando (nonprint) se introduce para que al ejecutar la instrucción el software la realice pero sin mostrar el resultado. Adicionalmente a cada serie de comandos se le puede agregar otros dos que permiten crear nuevos objetos y mostrar únicamente algún dato específico: **compute y display**, respectivamente; esto se hace con el fin de simplificar la ejecución de la prueba. Siguiendo el ejemplo, los comandos deben ser utilizados de la siguiente manera:

```
compute src = %rss , n=%nobs
display 'SRC' src
```

Compute está indicándole al software que nombre la Suma de Residuos al Cuadrado como 'src' y el número de observaciones 'n' con el fin de identificarlos más fácilmente. **Display** está indicando que se desea visualizar la palabra 'SRC' junto con el resultado de la Suma de Residuos. Los comandos fueron introducidos para las tres regresiones.

A su vez la instrucción **compute** se utiliza para construir el estadístico F en dos niveles: el primero unifica las SRC y el número de observaciones de las dos sub-muestras y el segundo construye la fórmula del estadístico. Por último se agrega el comando **CDF** (Función de Densidad Acumulada), junto con argumentos propios de la prueba F, en el cual se especifican los grados de libertad utilizados.

```
compute src12=src1+src2, n12=n1+n2
compute fstat=((src-src12)/5)/(src12/46)
cdf ftest fstat 5 46
```

Ilustración 37. Cálculo del estadístico F-Test de Chow-RATS

El resultado arroja el valor calculado de la F y su nivel de significancia para contrastar la prueba de hipótesis.

Para la prueba de CUSUM, en RATS se debe digitar una combinación de instrucciones que incluyen los comandos **compute, set** y **graph** para construir tanto la curva de los residuos como los límites de confianza. En primer lugar debe estimarse la regresión lineal adicionando en la parte inferior del comando la opción de que imprima los residuos,

```
linreg(print,define=R1) y / res
#constant x1 x2 x3 x4
print / res
```

Luego debe correrse la primera serie de instrucciones que le permitirán al usuario guardar ese vector de residuos obtenido como un archivo de Excel (debe recordarse la ubicación debido a que el archivo será utilizado enseguida).

```
open copy
copy(format=xls,org=obs) 1994:01 2007:04 res
```

Ilustración 38. Guardar residuos-RATS

Finalmente el usuario debe digitar la siguiente línea de código bajo la cual el software calcula los componentes de la prueba y realiza finalmente el gráfico requerido.

```
cal 1994 01 04
all 2007:04
open data
data(format=xls,org=col) / res
compute n=56
compute A=0.948
compute C0=0.032515
compute C0=0.04043
set cusum / = 0.0
compute cusum(1)=res(1)
do it=2,n
compute cusum(it)=cusum(it-1)+res(it)
end do it
statistics(noprint) res
set cusum 1 n = cusum(t)/sqrt(%variance)
set ls 1 n = A*sqrt(n-1)+2*A*(t-1)/sqrt(n-1)
set li 1 n = -ls(t)
set lqs 1 n = C0+float((t-1))/float((n-1))
set lqi 1 n = -C0+float((t-1))/float((n-1))
graph(header='CUSUM',subheader="") 3
# cusum 1 n
# ls 1 n 2
# li 1 n 2
```

Ilustración 39. Prueba CUSUM-RATS

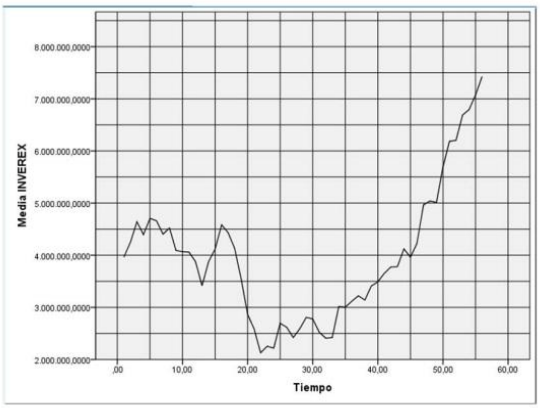
2.5.2.1.4. SPSS

SPSS no incluye explícitamente ningún test formal de los anteriormente explicados para verificar la existencia de cambio estructural. La única herramienta útil que podemos encontrar para éste efecto es el método gráfico.

Utilizando el *Generador de Gráficos* que se encuentra en la pestaña *Gráficos*, podemos visualizar el comportamiento que ha tenido la variable dependiente a través del tiempo[24]. Al abrir el asistente de gráficos debemos escoger el tipo que deseamos, en este caso Línea y desplazar las variables hacia los ejes respectivos.

A modo de ilustración se obtuve el siguiente resultado a partir del cual se puede inferir por el cambio de tendencia, la existencia de un cambio estructural en algunos puntos del tiempo; sin embargo no es posible llegar a una conclusión rigurosa.

[24] Variable incluida en este caso particular.



2.5.3. HIPÓTESIS DE ESPECIFICACIÓN ERRÓNEA

La especificación errónea del modelo puede hacer referencia a tres situaciones distintas:

- Omisión de variables relevantes.
- Inclusión de Variables no relevantes.
- Forma funcional no adecuada.

Cuando se omiten variables que pueden ser importantes, se obtienen por un lado, estimadores sesgados e inconsistentes y por otro lado la suma de residuos al cuadrado y la varianza de los errores será mucho mayor de lo que debería ser en realidad, haciendo que se pierda confiabilidad en la estimación. Además, en esos casos, el modelo puede aparentar cambios estructurales o problemas de autocorrelación, debido a cambios que se producen en la variable omitida.

En la segunda situación, la inclusión de variables irrelevantes genera estimadores consistentes pero insesgados; en este caso la varianza de los errores y de los estimadores aumenta haciendo que las pruebas de significancia individual para las variables pierdan confiabilidad.

La última situación a la que se asocia la especificación errónea es la escogencia de una forma funcional no adecuada para el modelo, lo cual produce las mismas consecuencias que en la primera situación explicada.

En los dos primeros casos no existe una prueba rigurosa que permita identificar la omisión o inclusión de variables relevantes. En el primer caso es importante revisar con detalle la

teoría económica para encontrar relaciones entre las variables y en el segundo deberá re-estimarse el modelo excluyendo las variables que no son significativas.

Para el caso de la especificación errónea por forma funcional inadecuada, la prueba utilizada para evaluar su presencia es el test RESET (REgression Epecification Error Test) de Ramsey, en la cual se comprueba si las combinaciones no lineales de los valores estimados ayudan a explicar la variable endógena; si estas combinaciones no lineales resultan explicando de algún modo significativo la variable dependiente, el modelo en consecuencia está mal especificado.

La aplicación del test se hace de la siguiente manera:

- Estimar la regresión de forma normal, calcular el coeficiente de determinación (R^2) y generar los valores estimados para Y.
- Estimar la regresión incluyendo además de las variables exógenas, la variable endógena elevada a diferentes potencias: Y^2, Y^3,…,Y^n. Calcular el coeficiente de determinación (R^2_1).
- Calcular el estadístico F mediante la ecuación:

$$F = \frac{\frac{(R_1^2 - R^2)}{(K_{H_o})}}{\frac{(R_1^2)}{(n-k)}}$$

- Finalmente, si el estadístico calculado es mayor al valor crítico de la distribución F, se concluye que el modelo está mal especificado; o si el valor-p es menor a α, se rechaza la hipótesis nula, de lo contrario no se rechaza.

2.5.3.1. APLICACIÓN EN SOFTWARE

2.5.3.1.1. Stata 11.0

La aplicación del test RESET de Ramsey en Stata es muy sencilla debido a que el software cuenta con un comando diseñado específicamente para probar errores de especificación en un modelo de regresión a través de dicha prueba.

El comando es **ovtest** y no tiene argumentos que lo complementen; sin embargo, ofrece al usuario la opción de digitarlo de la forma **ovtest, rhs**, lo cual indica que la prueba va a ser realizada elevando las variables independientes a diferentes potencias, y no la variable

dependiente como comúnmente se hace. El resultado obtenido se muestra en la ilustración 40.

```
. ovtest

Ramsey RESET test using powers of the fitted values of inverex
       Ho:  model has no omitted variables
                  F(3, 48) =      0.81
                  Prob > F =      0.4966
```

Ilustración 40. Prueba RESET de Ramsey-Stata

Allí se observa una descripción y especificación básica de la prueba[25], el valor del estadístico y su nivel de significancia. A partir de esos datos, el modelo de regresión utilizado estaría correctamente especificado.

2.5.3.1.2. R-project

Al igual que Stata, R cuenta con un código específico para la realización del test RESET de Ramsey, el cual es muy fácil de identificar ya que lleva el nombre de la prueba: **reset**. Se encuentra especificado de la siguiente forma:

```
reset(formula, power = 2:3, type = c("fitted", "regressor",
  "princomp"), data = list())
```
[26]

El primer y último argumento son los mismos vistos en aplicaciones anteriores de comandos del software: una fórmula que especifique el modelo a evaluar y el conjunto de datos del cual se extrae. Los dos argumentos del medio son propios de la prueba: *power* indica las potencias que van a ser incluidas escritas en forma de vector (#:#) y *type* permite elegir entre las variables que van a ser elevadas: la variable dependiente ajustada, todas las variables regresoras o el primer componente principal de la matriz de regresoras. Por defecto, la prueba se aplica con las potencias cuadrática y cúbica sobre la variable dependiente.

Para efectos de mantener los resultados obtenidos con Stata, el comando se utilizó como se ilustra.

[25] Por defecto, Stata realiza la prueba incluyendo desde la segunda hasta la cuarta potencia.

[26] Especificación tomada de la ayuda del software.

```
> reset(INVEREX~PIBR+E+INF+IR, power = 2:4, data = Base1)

        RESET test

data:  INVEREX ~ PIBR + E + INF + IR
RESET = 0.8062, df1 = 3, df2 = 48, p-value = 0.4966
```

Ilustración 41. Prueba RESET de Ramsey-R

En la salida se observa además del valor del estadístico F y el valor-p, los grados de libertad utilizados.

2.5.3.1.3. WinRATS 7.2

En este software es necesario programar el test de forma manual, es decir crear por separado las variables elevadas a las diferentes potencias, para luego correr la regresión incluyéndolas y aplicar la prueba F del test. Para crear las nuevas variables elevadas de forma rápida y sencilla se debe utilizar la función **prj fitted** la cual computa valores ajustados basados en la información de la última regresión efectuada. A continuación se deberán establecer las nuevas variables, elevándolas a las potencias de la forma **#, como lo muestra el proceso.

```
prj fitted
set fit2 = fitted**2
set fit3 = fitted**3
   ⋮
```

Como paso siguiente se debe volver a efectuar una regresión que incluya, además de las variables independientes, las nuevas variables. Por último, para realizar la prueba F que determina si el modelo está correctamente especificado, recurrimos a la función **exclude,** la cual calcula el estadístico de prueba bajo la hipótesis nula de que los coeficientes de las variables en la regresión son iguales a cero; esta función debe usarse siempre después de hacer una regresión.

```
exclude
# fit2 fit3 fit4
```

Ilustración 42. Estadístico F-Prueba RESET de Ramsey-RATS

Como resultado, se obtiene el valor del estadístico con su nivel de significancia para la prueba.

2.5.3.1.4. SPSS

SPSS no incluye ninguna prueba, test o estadístico riguroso para verificar la correcta o errónea especificación del modelo de regresión. El análisis de esto requiere en este caso una mayor atención del investigador a la hora de consultar bibliografía relacionada con el tema; igualmente podrá revisar el cumplimiento de las propiedades de los estimadores, lo cual en algunos casos puede mostrar evidencia de este problema.

2.5.4. HIPÓTESIS DE MULTICOLINEALIDAD

Como se mencionó anteriormente, uno de los supuestos importantes del modelo es la independencia entre las variables exógenas, lo cual estaría siendo violado por esta hipótesis. Existen dos tipos de multicolinealidad: la multicolinealidad Exacta y la Aproximada. La primera se denomina así debido a que es el caso en el cual se presenta la existencia de una combinación lineal exacta entre dos o más variables exógenas incluidas en el modelo; y la segunda se dice aproximada cuando existe una relación fuerte más no exacta entre dos o más variables.

La multicolinealidad es un fenómeno difícil de evitar debido a la estructura de la misma economía en la cual existe multiplicidad de interrelaciones. Ésta se produce por dos razones: en primer lugar por un error en la especificación del modelo en el cual no se observó la existencia de una identidad o causalidad que liga a las variables, y en segundo lugar cuando se trabaja con variables cualitativas, las cuales son representadas generalmente por variables ficticias o dummy. Este caso particular se conoce como la "Trampa de las variables ficticias", la cual consiste en la inclusión de una variable de este tipo por cada categoría o nivel existentes en el modelo, haciendo que se genere un problema de dependencia entre las variables exógenas.

Como todo lo visto anteriormente, la violación del supuesto de multicolinealidad tiene serias consecuencias sobre el modelo de regresión lineal. En el caso de la multicolinealidad Exacta, la principal consecuencia es que no se puede estimar los β debido a que no es posible calcular la inversa de la matriz X'X ya que su determinante es igual a cero. Por el lado de la multicolinealidad Aproximada se presentan varias situaciones: los estimadores pueden tener magnitudes no lógicas y/o signos distintos a lo esperado; las varianzas de los estimadores van a ser mayores de lo que deberían ser haciendo que se pierda confianza en las pruebas de inferencia estadística; los estimadores se vuelven sensibles cuando se añade

nueva información al modelo y por último puede presentarse una contradicción entre las pruebas de significancia individual, significancia global y el R2 del modelo, haciendo que no haya certeza sobre la viabilidad del modelo.

Para la detección del problema de multicolinealidad, al contrario de los demás supuestos del modelo de regresión, no se han establecido pruebas o contrastes estadísticos concretos que determinen con exactitud la existencia de relaciones fuertes entre las variables exógenas, sin embargo si existen una serie de reglas y posibles recetas que pueden proveer resultados precisos.

Coeficientes de Correlación

La primera forma y la más inmediata, consiste en encontrar los coeficientes de correlación simple entre las variables exógenas, para lo cual se debe encontrar la matriz de correlaciones de las variables. Este coeficiente se denota como $\mathbf{r}_{ij}$, halla la correlación existente entre pares de variables (X_i y X_j) y se encuentra acotado en el dominio de -1 a 1. La regla establece que si el coeficiente calculado es cercano a 0 las variables no se encuentran relacionadas, mientras que si es cercano a -1 o 1 (valores mayores a -0.8 y 0.8 respectivamente)[27] puede presentarse multicolinealidad debido a que las variables se encuentran muy relacionadas.

Regresiones Auxiliares

Otro método es efectuar regresiones auxiliares entre las variables exógenas, es decir tomar cada variable exógena como si fuera endógena y regresarla con las otras exógenas; el proceso debe repetirse para cada variable exógena del modelo. En cada regresión efectuada, se debe observar el coeficiente de determinación R^2 y si éste es mayor al R^2 de la regresión del modelo original, entonces es síntoma de multicolinealidad.

Índice Condición

La tercera forma de detectar posibles relaciones fuertes entre variables exógenas es a partir de la construcción del Índice de Condición, el cual es la raíz cuadrada del número condición.

$$IC = \sqrt{\frac{\lambda_{máx}}{\lambda_{mín}}}$$

[27] En algunos casos se puede ser flexible y establecer que a partir de 0.7 las variables se encuentran fuertemente relacionadas.

El número condición es el cociente entre el máximo valor propio y el mínimo valor propio de la matriz X'X. La regla práctica dice que si el Índice se encuentra entre 10 y 30, se presenta multicolinealidad moderada y si es mayor a 30, multicolinealidad severa.

Factores De Tolerancia y De Inflación De Varianza

Los factores de inflación de varianza (VIF) y de tolerancia (TOL) son ampliamente utilizados para la evaluación de la multicolinealidad. El VIF está definido como $\frac{1}{1-R_j^2}$, de donde se deduce que si el R^2 es muy alto (cercano a 1), la varianza de los estimadores se inflará en una gran proporción debido a la presencia de colinealidad entre las variables. Se tiene entonces, que cuando el VIF es mayor a 10, lo que equivale a decir que R^2=0.9, la variable está altamente correlacionada con otra u otras.

En cuanto al factor de tolerancia, éste está estrechamente relacionado con el VIF, siendo definido como $\frac{1}{VIF}$. De ésta forma, entre más cercano a 0 sea el valor de TOL, mayor será el grado de colinealidad de las variables, mientras que más cercano sea a 1 es evidencia de que no hay multicolinealidad (Gujarati, 2003) .

2.5.4.1. APLICACIÓN EN SOFTWARE

2.5.4.1.1. Stata 11.0

Stata permite al usuario detectar problemas de multicolinealidad mediante todos los métodos mencionados. Para todos posee un comando especial excepto para la realización de las regresiones auxiliares, las cuales se hacen como una regresión normal pero rotando las variables exógenas entre sí.

Como primera medida, para obtener los coeficientes de correlación entre las variables, se debe calcular la matriz de correlaciones (matriz simétrica) utilizando el comando **corr**[28] y poniendo como argumentos las variables explicativas,

[28] El investigador podrá observar también las correlaciones parciales y semi-parciales con la ayuda del comando **pcorr**.

```
. corr  pibr e inf ir
(obs=56)
```

	pibr	e	inf	ir
pibr	1.0000			
e	0.4976	1.0000		
inf	-0.7036	-0.8356	1.0000	
ir	-0.5233	-0.6228	0.6338	1.0000

Ilustración 43. Matriz de Correlaciones-Stata

Para recurrir a los demás métodos, Stata posee un comando que integra las demás medidas para detectar la colinealidad entre las variables: VIF, TOL y el IC. El comando, al igual que el utilizado para desarrollarla prueba de CUSUM, no se encuentra dentro de los comandos bases del software por lo que debe ser instalado a través del comando **findit**, el cual generará una ventana en la cual se podrá buscar el paquete y luego instar. El nombre del paquete y del comando es **collin** y debe utilizarse seguido de las variables del modelo que están siendo evaluadas.

```
. collin  pibr e inf ir
(obs=56)

  Collinearity Diagnostics
```

Variable	VIF	SQRT VIF	Tolerance	R-Squared
pibr	2.18	1.48	0.4584	0.5416
e	3.76	1.94	0.2661	0.7339
inf	5.27	2.30	0.1898	0.8102
ir	1.83	1.35	0.5464	0.4536
Mean VIF	3.26			

	Eigenval	Cond Index
1	4.6279	1.0000
2	0.2923	3.9793
3	0.0725	7.9889
4	0.0065	26.7531
5	0.0009	72.4525

```
Condition Number         72.4525
Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
Det(correlation matrix)    0.0788
```

Ilustración 44. Multicolinealidad-Stata

El resultado expresado en la ilustración 44 incluye, además de las medidas mencionadas, el VIF elevado al cuadrado y los valores propios de la matriz a partir de los cuales se calcula el Índice Condición.

Igualmente, si el usuario desea observar por separado los factores VIF y TOL, podrá hacerlo digitando únicamente el comando **estat vif**[29] sin necesidad de especificar las variables explicativas, las cuales son reconocidas automáticamente por el software.

Tras el cálculo de todas las medidas propuestas, se puede concluir que aparentemente la variable *INF* está medianamente relacionada de forma negativa con las demás variables exógenas del modelo por lo que se podría presentar problemas de multicolinealidad.

2.5.4.1.2. R-Project

R-project permite aplicar todos los métodos para medir la colinealidad de las variables exógenas. Adicionalmente, ofrece el gráfico de dispersión descrito en la sección de Estadística Descriptiva, con el cual puede observarse patrones en el comportamiento de las variables para concluir si presentan o no alguna relación fuerte.

Para comenzar con la matriz de correlaciones, el comando a utilizar es **cor()**, incluyendo en el paréntesis el nombre del conjunto de datos, en este caso la base de datos original sin la primera columna (variable dependiente).

```
> cor(datosrh)
           PIBR          E        INF         IR
PIBR  1.0000000  0.4976327 -0.7036317 -0.5232889
E     0.4976327  1.0000000 -0.8356125 -0.6227744
INF  -0.7036317 -0.8356125  1.0000000  0.6337746
IR   -0.5232889 -0.6227744  0.6337746  1.0000000
```

Ilustración 45. Matriz de Correlaciones-R

Los coeficientes de correlación en R pueden ser hallados a partir de un test para calcular el grado de asociación entre pares de variables; el comando para aplicarlo es **cor.test(X,Y)**[30] en donde x y y son dos vectores numéricos con la misma longitud que representan las variables a evaluar. El test hace parte de un paquete llamado *stats*, el cual debe ser previamente instalado. A manera de ilustración se calculó el coeficiente de relación entre las variables PIBR y E, obteniéndose el mismo resultado:

[29] En versiones anteriores del software el comando recibía el nombre de **vif**.

[30] Por defecto la correlación es calculada con el coeficiente de Pearson, sin embargo el test ofrece la opción de cambiarlo por el de Kendall o Spearman.

```
> cor.test(PIBR,E)

        Pearson's product-moment correlation

data:  PIBR and E
t = 4.2159, df = 54, p-value = 9.534e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2700641 0.6725455
sample estimates:
      cor
0.4976327
```

Ilustración 46. Coeficientes de correlación-R

En cuanto al factor VIF, la función para calcularlo hace parte del paquete *HH* y se identifica con el mismo nombre del factor. En este caso, la función se encuentra especificada así: **vif(x,…)**; en donde x puede representar una fórmula, un objeto tipo data.frame o un objeto tipo lm. En este caso se aplicó como si x fuera un data.frame, para lo que fue necesario convertir primero la matriz de las variables exógenas en un objeto de este tipo, como se ilustra a continuación,

```
> frame1=data.frame(datosrh)
> vif(frame1)
    PIBR        E      INF       IR
2.181426 3.758532 5.267887 1.830241
```

Ilustración 47. VIF-R

En cuanto al Índice de Condición la función también hace parte de un paquete que debe ser instalado: *perturb*. El comando que ejecuta la instrucción es **colldiag()** y al igual que el comando **vif** necesita que el objeto a evaluar sea de tipo data.frame. El resultado obtenido es el Índice junto con las proporciones de varianza descompuestas por cada variable.

```
> colldiag(frame1)
Condition
Index   Variance Decomposition Proportions
          intercept PIBR  E     INF   IR
1   1.000 0.000     0.000 0.000 0.002 0.004
2   3.979 0.000     0.003 0.002 0.074 0.055
3   7.989 0.000     0.000 0.000 0.227 0.835
4  26.753 0.003     0.530 0.213 0.000 0.000
5  72.453 0.997     0.466 0.785 0.697 0.105
```

Ilustración 48. Índice de Condición-R

2.5.4.1.3. WinRATS 7.2

RATS maneja una matriz llamada Covariance\Correlation Matrix, la cual es un resultado combinado de las covarianza y correlaciones existentes entre las variables, la primera siendo representada por la diagonal y la parte debajo de ésta y las segundas encontradas en la parte superior.

Esta matriz puede obtenerse de dos formas: mediante el código VCV utilizando como carta suplementaria las variables de interés (en este caso las explicativas), o recurriendo a la pestaña ***Statistics > Covariance Matrix,*** en donde aparecerá una ventana en la cual se debe elegir las variables de interés.

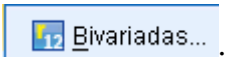
Ilustración 49. Matriz de Covarianzas y Correlaciones-RATS

Al confirmar las variables deberá presionar *OK* para ejecutar la instrucción. Notará que en el input se genera el mismo código VCV y en el output aparece la matriz no simétrica.

Adicionalmente, para evaluar el supuesto podrá realizar las regresiones auxiliares utilizando el comando de regresión.

2.5.4.1.4. SPSS

Para la evaluación del supuesto de multicolinealidad en el modelo de regresión lineal, SPSS cuenta con las mismas herramientas ofrecidas por los demás software. A diferencia de estos, su implementación no requiere de ningún comando sino del seguimiento de unos cuantos pasos muy sencillos.

Por un lado, para obtener la matriz de correlaciones es necesario dirigirse a la pestaña *Analizar* ubicada en el menú superior, allí seleccionar el sub-menú *Correlaciones* y por último el ícono 

. Como parte del proceso se abrirá una nueva ventana en donde deben ser seleccionadas tanto las variables para las cuales se va a medir el grado de

correlación como el Coeficiente de Correlación a utilizar, el cual generalmente es Pearson. A su vez, se presenta una opción para que el software resalte las correlaciones que considere significativas, herramienta muy útil para guiar al usuario en su análisis. Luego de llenar los campos correspondientes y seleccionar ACEPTAR, la matriz de correlaciones generada es la siguiente:

Correlaciones

		PIBR	E	INF	IR
PIBR	Correlación de Pearson	1	,498**	-,704**	-,523**
	Sig. (bilateral)		,000	,000	,000
	N	56	56	56	56
E	Correlación de Pearson	,498**	1	-,836**	-,623**
	Sig. (bilateral)	,000		,000	,000
	N	56	56	56	56
INF	Correlación de Pearson	-,704**	-,836**	1	,634**
	Sig. (bilateral)	,000	,000		,000
	N	56	56	56	56
IR	Correlación de Pearson	-,523**	-,623**	,634**	1
	Sig. (bilateral)	,000	,000	,000	
	N	56	56	56	56

**. La correlación es significativa al nivel 0,01 (bilateral).

Ilustración 50. Matriz de Correlaciones-SPSS

Por otro lado, SPSS resume igualmente en una tabla las demás medidas de colinealidad. Ésta opción aparece en la ventana utilizada para desarrollar la regresión en el ícono de ***Estadísticos***. En la parte derecha de la ventana de *Estadísticos* aparece la opción ☑ Diagnósticos de colinealidad, la cual luego de haber sido seleccionado y estimada la regresión, genera como resultado las siguientes tablas:

Estadísticos de colinealidad	
Tolerancia	FIV
,458	2,181
,266	3,759
,190	5,268
,546	1,830

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalores	Índice de condición	Proporciones de la varianza				
				(Constante)	PIBR	E	INF	IR
1	1	4,628	1,000	,00	,00	,00	,00	,00
	2	,292	3,979	,00	,00	,00	,07	,06
	3	,073	7,989	,00	,00	,00	,23	,84
	4	,006	26,753	,00	,53	,21	,00	,00
	5	,001	72,453	1,00	,47	,78	,70	,11

a. Variable dependiente: INVEREX

Ilustración 51. Diagnósticos de Colinealidad-SPSS

La tabla de la derecha aparece adherida a la tabla de resultados de la regresión, e incluye los factores VIF y TOL. La tabla de la izquierda muestra en detalle el Índice de Condición

junto con las proporciones que añade cada variable por separado a la construcción de la varianza.

2.5.5. COMPARACIÓN DE LAS PRUEBAS DE ESTRUCTURA EN LOS SOFTWARE

A partir de la validación de los supuestos acerca de la estructura del modelo de regresión lineal que se llevó a cabo previamente utilizando todos los software como medio de aplicación, fue posible observar las virtudes, deficiencias y dificultades de cada uno de ellos en cuanto a las herramientas ofrecidas para el contraste de estos supuestos. Los resultados por cada supuesto se ilustran en la tabla I.

SOFTWARE	CAMBIO ESTRUCTURAL	ESPECIFICACIÓN ERRÓNEA	MULTICOLINEALIDAD
Stata	Aplicación parcial de los contrastes.	Desarrollo preciso y automático del contraste. Variedad de opciones para su aplicación.	Gran variedad de métodos de contraste. Código de aplicación simple.
R-Project	Amplia aplicación de contrastes. Variedad de alternativas para la estimación de cada prueba.	Desarrollo preciso y automático del contraste. Variedad de opciones para su aplicación.	Análisis completo de presencia de multicolinealidad.
WinRATS	Necesidad de líneas de código de difícil comprensión debido a la inexistencia de comandos exactos.	Programación manual de la prueba de contraste. Necesidad de conocimiento sobre la teoría.	Herramientas que producen un análisis parcial. Metodología de estimación confusa.
SPSS	Inexistencia de contraste	Inexistencia de contraste	Herramientas poderosas para el contraste del supuesto. Obtención fácil e intuitiva de resultados.

Tabla I. Comparación de herramientas ofrecidas por los software en cuento al contraste de supuestos de la estructura del modelo.

2.6. SUPUESTOS SOBRE LOS RESIDUOS

2.6.1. SUPUESTO DE HOMOSCEDASTICIDAD

Para la estimación del modelo de regresión lineal es muy importante el supuesto de homoscedasticidad, bajo el cual la varianza de los errores permanece constante a lo largo de toda la muestra. Sin embargo, y por diversas causas, la varianza del término de perturbación aleatoria puede ser variable en el tiempo presentándose problemas de HETEROSCEDASTICIDAD[31]:

$$E(u_i^2) = \sigma_i^2$$

Ahora, el subíndice i indica que la varianza de los residuos ya no es constante.

Existen varias razones por las cuales las varianzas pueden ser variables (Gujarati, 2003):

- Corrección de errores de comportamiento por parte de las personas conforme pasa el tiempo y aumenta su aprendizaje.
- Mejoras en las técnicas de recolección y procesamiento de la información.
- Presencia de factores atípicos (observaciones con información muy diferente en relación a las demás observaciones).
- Especificación errónea del modelo de regresión, especialmente omisión de variables relevantes.
- Distribución asimétrica de las variables explicativas del modelo.
- Transformación incorrecta de las variables.

La violación al supuesto de homoscedasticidad, produce un cambio importante en la estimación del modelo por medio del método de MCO: el estimador ya no va a ser el más eficiente. Por un lado, la linealidad, insesgamiento y consistencia de $\hat{\beta}$ se mantienen debido a que la heteroscedasticidad no tiene ningún efecto sobre su determinación. Sin embargo, la varianza del estimador dejará de ser mínima debido a que la matriz de Varianzas-Covarianzas ya no es escalar, es decir

$$Var.Cov(U) = E(UU') = \sigma_u^2 \Omega \qquad con\ \Omega \neq \mathrm{I}$$

Esto significa que la varianza del estimador estará dada ahora por la siguiente ecuación:

[31] Éste problema se presenta con más frecuencia en datos de corte transversal, es menos probable en datos temporales.

$$Var(\hat{\beta}) = (X'X)^{-1}\sigma_u^2 \Omega\, X(X'X)^{-1}$$

Ilustración 52. Varianza de los estimadores en presencia de heteroscedasticidad

Adicionalmente, si se estima el modelo por MCO ignorando o desconociendo la presencia de heteroscedasticidad, las pruebas de significancia van a perder confiabilidad y se pueden producir conclusiones erróneas.

Teniendo en cuenta esto, aparece otro método de estimación que soluciona el problema de mínima varianza, siendo capaz de incluir la "información" contenida en la variabilidad de las variables, dándole a cada una de ellas una ponderación acorde a ésta: los estimadores de Mínimos Cuadrados Generalizados (MCG).

Para la construcción de estos estimadores es necesario partir del nuevo argumento de la matriz de varianzas y covarianzas Ω; al ser ésta una matriz definida positiva se puede descomponer en dos matrices simétricas, invertibles y no singulares; de la forma $\Omega = PP'$.

Suponiendo que la varianza heteroscedastica es proporcional a una constante σ^2 por una variable λ_i, la matriz Ω se construye con los elementos de la variable que es proporcional a la varianza de los errores ubicada en la diagonal y los demás valores 0. Siguiendo esto, la matriz P se conoce como la matriz inversa raíz cuadrada, construyéndose con el inverso de la raíz de la variable en la diagonal.

$$\Omega = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{pmatrix} \qquad P = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{\lambda_n}} \end{pmatrix}$$

Así los estimadores de Mínimos Cuadrados Generalizados están definidos por la ecuación,

$$\hat{\beta}_{MCG} = (\mathrm{X}'\Omega^{-1}\mathrm{X})^{-1}\,\mathrm{X}'\Omega^{-1}\,\mathrm{Y}$$

y son los estimadores MELI ante la presencia de heteroscedasticidad.

La violación al supuesto de homoscedasticidad se puede abordar por otro lado: se hace una transformación a todas las variables del modelo, pre-multiplicándolas por la matriz P y finalmente se estima el modelo resultante por el método de MCO. Así se concluye que la aplicación del método de MCO sobre variables transformadas que satisfacen los supuestos

de los estimadores de MCO produce el mismo resultado que aplicar el método de MCG al modelo original.

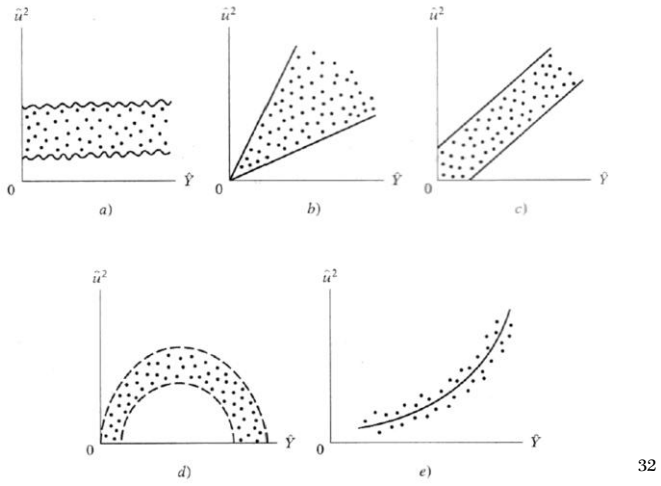
Al igual que la multicolinealidad, para detectar la heteroscedasticidad no hay reglas exactas, sin embargo existe una serie de reglas y métodos formales e informales que se utilizan con frecuencia.

Método Gráfico

Si no se evidencia en el modelo que existen posibles causas de heteroscedasticidad, se puede estimar una regresión normal por el método de MCO bajo el supuesto de homoscedasticidad para luego observar el comportamiento gráfico que tienen los términos de error. En primer lugar se debe estimar la regresión y calcular el vector de residuos, para luego generar dos tipos de gráficos:

a) Los residuos elevados al cuadrado $\hat{u}_i^2$ frente a la variable endógena estimada $\widehat{Y}_i$.
b) Los residuos elevados al cuadrado frente a cada una de las variables exógenas X_i.

La idea general es ver a partir de los gráficos si por un lado, el valor estimado de Y presenta algún patrón de relación con el término de error al cuadrado, o si el termino de error está relacionado de alguna forma (lineal, cuadrática, etc.) con alguna variable explicativa. En cualquiera de los dos casos se pueden obtener los siguientes resultados:



[32]

Ilustración 53. Evaluación heteroscedasticidad por método gráfico

[32] Ilustración tomada del libro *Eocnometría* de Damodar Gujarati (2003). Página 387.

En el gráfico a) no se observa ninguna relación definida, mientras que en los demás aparece claramente un patrón de relación, como en el gráfico c) el cual sugiere una relación lineal.

Prueba de Park

Ésta prueba se conoce como la formalización del método gráfico. Parte del supuesto de que la varianza heteroscedastica es proporcional a una constante σ^2 por una variable exógena de la forma X_i^{β} por el antilogaritmo (exponencial) de los residuos, o escrito de forma matemática

$$\ln(\sigma_i^2) = \ln(\sigma^2) + \beta\, lnX_i + e_i$$

Dado que generalmente el término σ_i^2 no se conoce, Park sugiere trabajar con los residuos al cuadrado como una aproximación, obteniendo el siguiente modelo de regresión:

$$\ln(\widehat{U}_i^2) = \alpha + \beta \ln(X_i) + e_i$$

Finalmente debe aplicarse una prueba de significancia al β; si éste resulta ser significativo, hay presencia de heteroscedasticidad, mientras que si no es significativo se cumple el supuesto de homoscedasticidad.

Prueba de Glejser

Ésta prueba es en esencia muy similar a la prueba de Park pero utiliza formas funcionales diferentes. Luego de haber calculado los residuos a partir de la estimación del modelo por MCO, sugiere hacer la segunda regresión sobre el valor absoluto de los residuos, así

$$|\widehat{U}_i| = \alpha + \beta\, X_i^h + e_i$$

Donde h toma los valores de { -1, 1, -½ , ½}.

Prueba Golfeld-Quand

La prueba supone que la varianza heteroscedástica está positivamente relacionada con una de las variables exógenas del modelo, elevada al cuadrado, es decir $\sigma_i^2 = \sigma^2\, X_i^2$. Para evaluar el supuesto, deben seguirse los siguientes pasos:

1. Ordenar las observaciones de forma ascendente según los valores de la variable exógena seleccionada.

2. Eliminar P observaciones centrales[33], donde P puede ser cercano a $\frac{n}{4}$. (Buscar que las sub-muestras restantes tengan el mismo tamaño).
3. Hacer regresiones por separado a cada una de las sub-muestras comenzando con el grupo de valores más pequeños de la variable X seleccionada (grupo de menor varianza). Obtener respectivamente las SRC.
4. Aplicar una prueba de hipótesis calculando el siguiente estadístico:

$$F_{CALC} = \frac{\frac{SRC_2}{gl_2}}{\frac{SRC_1}{gl_1}} \qquad \sim \qquad F_{TABLA}(gl_1, gl_2)$$

Donde $gl_i = N_i - k$. Si el estadístico calculado es mayor al estadístico de la tabla, se rechaza la hipótesis nula y se dice que hay heteroscedasticidad; de lo contrario no; o si el valor-p es menor al valor α se rechaza la hipótesis nula y si es mayor no se rechaza.

Prueba de White

La aplicación de esta prueba es muy sencilla, solo es necesario seguir tres pasos:

1. Estimar el modelo de regresión original y calcular el vector de errores.
2. Hacer una regresión auxiliar del vector de residuos al cuadrado contra las variables exógenas del modelo original, sus valores elevados al cuadrado y los productos cruzados de las explicativas. Obtener el R_j^2 de esa regresión.
3. Bajo la hipótesis nula de homoscedasticidad, calcular el estadístico Ji-Cuadrado

$$X^2_{CALC} = n\, R_j^2 \qquad \sim \qquad X^2_{TABLA}\ (K_j - 1)gl$$

La prueba se contrasta como se ha hecho anteriormente.

Prueba Breush-Pagan.Godfrey

Esta prueba de utiliza para evitar algunas limitaciones de la aplicación de Golfeld-Quand.
El procedimiento a seguir es:

1. Estimar la regresión por MCO y obtener el vector de residuos.
2. Calcular $\hat{\sigma}^2 = \frac{\sum \hat{U}_i^2}{n}$, siendo este el estimador de máxima verosimilitud de σ^2.
3. Construir unas variables P dividiendo el vector de residuos al cuadrado en $\hat{\sigma}^2$.
4. Hacer una regresión de los P obtenidos contra las variables explicativas.

[33] Esto se hace con el fin de crear un mayor contraste entre las sub-muestras.

5. Obtener la Suma Explicada de Cuadrados definida como $\Theta = \frac{1}{2}(SRC)$ y contrastarla con una Ji-Cuadrado con (k-1) grados de libertad.

El contraste de la prueba se hace bajo la hipótesis nula de homoscedasticidad.

2.6.1.2. APLICACIÓN EN SOFTWARE

2.6.1.2.1. Stata 11.0

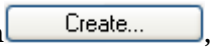
Stata cuenta con la mayoría de pruebas mencionadas anteriormente para evaluar el supuesto de homoscedasticidad. Como se vio, el primer método exploratorio para observar que la varianza de los residuos no es constante a lo largo de la muestra es realizando gráficos de dispersión de los residuos al cuadrado contra la variable endógena estimada o las variables exógenas.

En primera instancia es necesario obtener tanto la variable endógena estimada como los residuos para luego elevarlos al cuadrado. Para calcular predicciones de valores a partir de la regresión debe usarse el comando **predict**, el cual va acompañado del nombre de la nueva variable y otros argumentos según lo deseado: para obtener valores estimados de alguna variable se acompaña con **_hat** y para obtener residuos con **resid**, así:

```
. predict  inverex1_hat          . predict residual, resid
```

Para crear o cambiar el contenido de una variable se utiliza el comando **gen**, colocando en primer lugar el nuevo nombre de la variable en segundo la transformación que se vaya a llevar a cabo,

```
. gen res2 = residual^2
```

Teniendo esto, el gráfico puede realizarse de dos formas: a partir del comando **twoway()** en el cual hay que definir qué tipo de gráfica y variables se desea graficar o acudiendo a la pestaña *Graphics > Twoway graph (scatter, line, etc.)*, la cual produce le mismo resultado. Luego de haber seleccionado la opción aparecerá una ventana en la cual se debe definir un gráfico nuevo por medio del botón 

, y ahí, otra ventana en la cual, al igual que con el comando se debe definir el tipo de gráfica y escoger las variables.

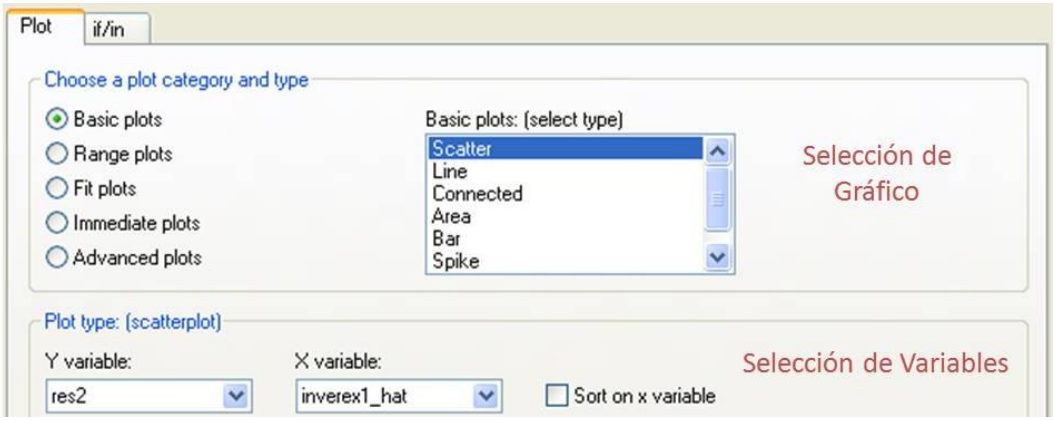
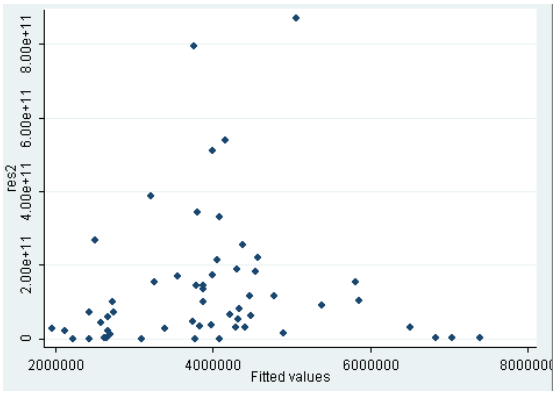


Ilustración 54. Generación de gráficos-Stata

Luego de validar la selección, en la ventana anterior aparecerá el nuevo gráfico creado, el cual se visualizará después de seleccionar OK.



Prueba Breusch-Pagan-Godfrey

Para aplicar esta prueba sobre la regresión en Stata, se tiene el comando **hettest**. No necesita ningún argumento ni especificación para ser ejecutado.

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of inverex

         chi2(1)      =     0.18
         Prob > chi2  =   0.6705
```

Ilustración 55. Prueba Breusch-Pagan-Stata

El resultado muestra una serie de detalles sobre la prueba como la hipótesis nula y la variable utilizada para su construcción, además del valor crítico del estadístico Ji-Cuadrado y el valor-p resultante.

Prueba de White

En el software, la ejecución de esta prueba viene acompañada de otra que realiza una prueba de la matriz de información[34] para el modelo, descomponiendo el estadístico en tests de heteroscedasticidad, simetría y kurtosis debido a Cameron y Trivedi. El resultado de la prueba de White concuerda con la primera fila de la matriz de información.

```
. estat imtest, white

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

         chi2(14)     =     24.15
         Prob > chi2  =     0.0440

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	24.15	14	0.0440
Skewness	3.14	4	0.5346
Kurtosis	.	1	.
Total	.	19	.

Ilustración 56. Prueba de White-Stata

2.6.1.2.2. R-Project

Como se hizo en el punto anterior, para recurrir al método gráfico es necesario obtener primero los residuos y los valores predichos de la variable explicada. Para cada uno de estos fines existe un comando a seguir: **resid()** y **fitted()** respectivamente, en donde dentro del paréntesis debe ir el objeto del cual se van a obtener dichos valores, en este caso el modelo de regresión lineal. Para facilitar la construcción del gráfico, se crearon dos objetos nuevos que incluyeran los resultados:

```
> res <- resid(multiple)
> res2 <- res^2
> fit <- fitted(multiple)
```

[34] Propuesto por White

A continuación se puede proceder a la obtención del gráfico, para lo cual se utiliza **plot()**, incluyendo dentro del paréntesis las variables a graficar, colocando primero la variable ubicada en el eje X y luego la variable ubicada en el eje Y,

```
> plot(fit,res2)
```

Prueba Breusch-Pagan-Godfrey

Para poder aplicar esta prueba, es necesario primero instalar el paquete del software que permite su ejecución, identificado con el nombre de **car**. Luego de instalar y llamar al paquete, el usuario debe digitar el comando **ncvTest()**, el cual da la facilidad de colocar en el paréntesis la información a evaluar de dos maneras diferentes: una fórmula que identifique simbólicamente el modelo a evaluar o el nombre de un objeto del tipo "lm". A modo de ilustración, se presenta a continuación el resultado obtenido:

```
> ncvTest(multiple)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1810193    Df = 1     p = 0.6704987
```

Ilustración 57. Prueba Breusch-Pagan-R

Este ilustra los grados de libertad empleados, el valor del estadístico y su valor-p.

Adicionalmente, el software presenta otro comando para ejecutar la prueba, para el cual debe instalarse el paquete **lmtest**. El comando esta caracterizado por contener las iniciales del nombre de la prueba: **bptest()**, incluyendo en el paréntesis los mismo argumentos que en la otra prueba. Sin embargo, esta prueba difiere a la anterior debido a que hace el cálculo del estadístico utilizando las variables explicativas y no los valores pronosticados de la explicada.

2.6.1.2.3. WinRATS 7.2

WinRATS permite aplicar las mismas pruebas formales que los demás software para la detección del problema de heteroscedasticidad. En primer lugar se encuentra la prueba de White, la cual necesita el cálculo previo de los residuos al cuadrado, los valores e las variables explicativas elevadas al cuadrado y sus productos cruzados para poder efectuar la regresión auxiliar.

Para cada uno de estos valores debe crearse una nueva variable en la cual se ejecute las operaciones anteriormente dichas, es decir que se multipliquen tanto las variables explicativas por sí mismas como entre ellas, como se ilustra:

```
set res2 = res*res
set x11 = x1*x1
set x1x2 = x1*x2
set x1x3 = x1*x3
```

El proceso debe repetirse hasta obtenerse valores para cada una de las variables y sus posibles combinaciones. Luego debe realizarse la regresión utilizando como variable explicada el vector de residuos al cuadrado y como explicativas todas las nuevas variables construidas,

```
linreg res2
# constant x1 x2 x3 x4 x11 x22 x33 x44 x1x2 x1x3 x1x4 x2x3 x2x4
```

Finalmente para contrastar la prueba de hipótesis, debe construirse el estadístico Ji-Cuadrado tal y como lo describe la prueba: el número de observaciones multiplicado por el R^2 de la regresión auxiliar; para lo cual se utiliza el comando **compute** y las instrucciones **%nobs** y **%rsquared** para obtener fácilmente los valores necesarios. Igualmente se debe aplicar la prueba de contraste del estadístico calculado con el valor del estadístico Ji-Cuadrado con (K_j – 1) grados de libertad. Todo esto se resume en:

```
display 'test de white'
compute chistat=%nobs*%rsquared
cdf chisqr chistat 13
```

Ilustración 58. Prueba de White-RATS

Prueba Breusch-Pagan-Godfrey

Esta prueba es calculada en el software de forma diferente a su descripción original hecha anteriormente, razón por la cual los resultados pueden cambiar en comparación con los de los demás software[35]. R-Project la calcula a partir del vector de residuos al cuadrado y el R^2 ajustado del modelo. Para su ejecución primero debe calcularse el vector de residuos al cuadrado, para luego efectuar una regresión auxiliar entre éste y las variables explicativas del modelo original. Finalmente debe realizarse el contraste del estadístico Ji-Cuadrado con el R^2 ajustado del modelo, el cual se obtiene mediante la instrucción **%trsquared**. Todo el proceso se ilustra a continuación:

[35] Procedimiento tomado de Estima. http://www.estima.com/textbooks/verp092.prg

```
set p = %resids**2
linreg p
# constant x1 x2 x3 x4
cdf(title="Breusch-Pagan Test") chisqr %trsquared 4

Breusch-Pagan Test
Chi-Squared(4)=     13.152167 with Significance Level 0.01055572
```

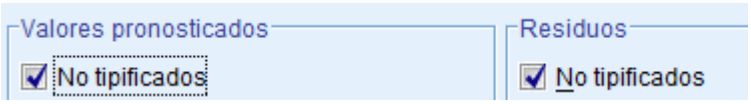
Ilustración 59. Prueba Breusch-Pagan-RATS

El resultado presenta el valor del estadístico Ji-Cuadrado utilizando como grados de libertad el número de variables exógenas del modelo y su nivel de significancia.

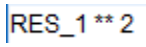
2.6.1.2.4. SPSS

Para contrastar el supuesto de homoscedasticidad SPSS incluye las mismas pruebas que los demás software, sin embargo su aplicación es más compleja y requiere un mayor trabajo.

Para recurrir al método gráfico, deben crearse nuevas variables que contengan los valores de la variable explicada predicha y los residuos. Estos valores se obtienen durante el proceso de estimación de la regresión lineal; en la ventana utilizada para este objeto aparece la opción *Guardar,* la cual abre una ventana en donde se deben seleccionar los Valores Pronosticados y los Residuos No Tipificados.

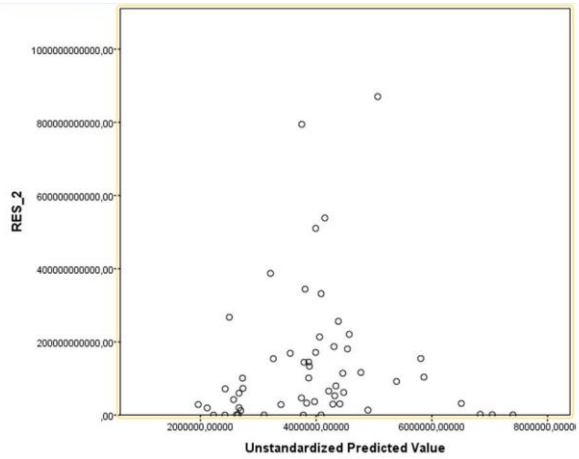


Luego de continuar y aplicar la regresión, dos nuevas variables aparecerán en la Vista de Datos: PRE_1 y RES_1. Para continuar la aplicación del método, es necesario elevar esos residuos al cuadrado, lo que se hace a través de las opciones *Transformar > Calcular Variable…*. Allí deberá seleccionar el nombre de la nueva variable (Variable de destino) y elevar los residuos al cuadrado (en Expresión Numérica) tal como se ilustra

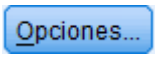


Luego de validar el cálculo, la nueva variable aparecerá igualmente en la Vista de Datos.

Para generar el gráfico, se recurre al mismo proceso empleado en la sección dedicada a la aplicación del test de Chow, con la diferencia de que aquí se desea graficar un diagrama de dispersión con la variable de residuos al cuadrado en el eje vertical y la variable explicativa predicha en el eje horizontal, obteniendo el siguiente resultado:

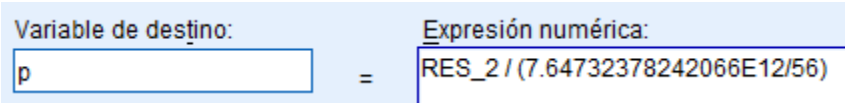


Prueba Breush-Pagan.Godfrey[36]

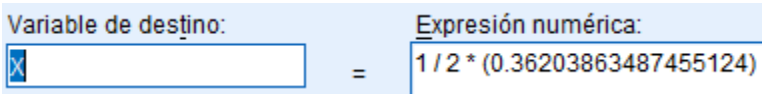
Ésta prueba necesita un proceso similar de construcción para su aplicación, para lo cual se van a utilizar las nuevas variables anteriormente generadas. El primer paso a seguir es calcular la Suma de Residuos al Cuadrado lo cual se hace a través de las opciones *Analizar > Estadísticos Descriptivos > Descriptivos.* Allí debe seleccionarse la variable que contiene el vector de residuos al cuadrado y dirigirse al ícono 

para seleccionar la opción de Suma.

Luego de tener ese resultado, debe calcularse una nueva variable que está definida por el vector de residuos al cuadrado dividido por la suma de residuos al cuadrado dividido a su vez por el número de observaciones; como ilustra la figura



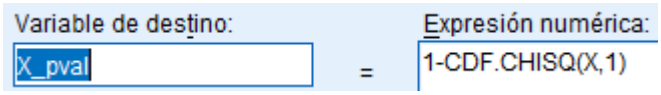
Luego, se debe estimar una regresión tomando p como la variable explicada y la variable endógena predicha como variable explicativa. A partir de esa regresión, se obtiene un nuevo valor para la suma de cuadrados[37], el cual se utilizará en el cálculo del estadístico de Breusch-Pagan: $\Theta = \frac{1}{2}(SRC)$, operación realizada por medio del cálculo de variables anteriormente utilizado,



[36] Proceso tomado de Introduction to SPSS.
http://www.kellogg.northwestern.edu/kis/tek/ongoing/Materials/Introduction2SPSS.pdf

[37] Valor ubicado en la tabla ANOVA que resulta de la regresión.

Cálculo a partir del cual se obtiene un valor X de 0,18101931743728. El paso final es contrastar esa X con una Ji-Cuadrado (obteniendo el valor-p para evaluar la prueba de hipótesis), utilizando la fórmula CDF.CHISQ[38] en la ventana de cálculo de variables:



Siguiendo éstos pasos el valor-p obtenido sobre el cual se va a concluir es 0,6704; resultado igual al obtenido con los demás software.

Prueba de White

La elaboración de la prueba de White no es tan sencilla en SPSS como lo es en los demás software, los cuales la incluyen de forma predeterminada en sus códigos. En éste, para la realización de la prueba se necesita utilizar una macro para programarla y construirla, un procedimiento que necesita conocimientos más avanzados por parte de los usuarios del software, razón por la cual no se considerará en éste caso.

A lo largo de todos los software se observó que los resultados de las pruebas varían de acuerdo a las diferentes metodologías de estimación utilizadas por defecto por cada uno de ellos. Cabe resaltar que la prueba Breusch-Pagan, una de las más importantes para la evaluación de este supuesto, indica que la varianza de los errores se mantiene constante a lo largo de toda la muestra; sin embargo los demás resultados son muy variados. Por lo tanto llegar a conclusiones en el ejemplo acerca de la presencia de heteroscedasticidad u homoscedasticidad no puede hacerse de modo preciso.

2.6.2. SUPUESTO DE NO AUTOCORRELACIÓN

En el modelo de regresión lineal, el supuesto de no autocorrelación implica que no existe correlación entre los términos de error de un período y el de otro u otros períodos, es decir

$$E(U_i U_j) = 0 \qquad i \neq j$$

En éste modelo, el problema que se presenta con mayor frecuencia es la autocorrelación de primer orden, en la cual los residuos están estrictamente relacionados con los residuos del periodo anterior lo cual se ilustra de la siguiente forma

$$U_t = \rho U_{t-1} + \varepsilon_t$$

[38] La fórmula se encuentra en la parte inferior derecha de la ventana

Sin embargo en los datos temporales puede presentarse autocorrelación hasta de orden p, en donde p es el número de períodos del denominado período autorregresivo.

Siguiendo la idea del problema de heteroscedasticidad, la matriz de varianzas y covarianzas va a seguir estando definida por la ecuación,

$$Var.Cov(U) = E(UU') = \sigma_u^2 \Omega \qquad con\ \Omega \neq I$$

Sin embargo ahora el problema se va a presentar en la información que se encuentra por fuera de la diagonal, ya que la covarianza de orden k de los errores va a estar dada por,

$$Cov(U_t, U_{t-k}) = \rho^k \sigma_u^2$$

donde ρ se conoce como el coeficiente de autocorrelación y se caracteriza por tener las mismas propiedades de un coeficiente de correlación ($-1 < \rho < 1$).

Las consecuencias de estimar un modelo por MCO en presencia de autocorrelación son las mismas que se presentan cuando de estima el modelo en presencia de heteroscedasticidad, es decir que los estimadores siguen siendo lineales e insesgados pero no van a tener varianza mínima por lo que no serán estimadores MELI. Por otro lado, los estimadores MCG si lo serán y aparecerán igualmente como una solución a ese problema.

La autocorrelación como se mencionó anteriormente es un problema que se presenta con mayor frecuencia en datos temporales y puede darse en la mayoría de los casos por la inercia propia de los acontecimientos económicos que hacen que lo que suceda hoy dependa de lo que sucedió en el pasado. Adicionalmente el problema puede presentarse por una especificación errónea del modelo, por un cambio estructural no considerado o por ignorar la presencia de variables endógenas rezagadas como variables explicativas del modelo.

Dentro de los métodos para detectar la presencia de autocorrelación, el más destacado y utilizado para detectar autocorrelación de primer orden es el llamado método Durbin Watson. Igualmente, para detectar autocorrelación de orden superior existen varios métodos como la prueba de Breusch-Godfrey o multiplicadores de Lagrange, la prueba de Box-Pierce, la prueba de Wallis y la prueba de Ljung-Box; las cuales no serán abordadas por motivos de consideración únicamente de autocorrelación de primer orden.

Método Gráfico

Como se ha visto en la comprobación de todos los supuestos del modelo de regresión lineal, el método gráfico resulta muy útil en la mayoría de los casos como una aproximación a la detección de diversos problemas, sin embargo no puede establecerse

como un método totalmente confiable debido a su subjetividad, por lo que siempre se encuentra soportado por otra serie de pruebas más formales.

En este caso, se desea observar si el comportamiento de los residuos de un período depende del comportamiento pasado, para lo cual se hará un gráfico de los residuos contra el tiempo. A partir de éste se podrá observar tres síntomas de autocorrelación: autocorrelación positiva (cuando el vector de residuos no cambia mucho de signos), autocorrelación negativa (cuando el vector de residuos cambia de signo constantemente) y no autocorrelación (cuando se observa mucha aleatoriedad en el comportamiento del vector de residuos).

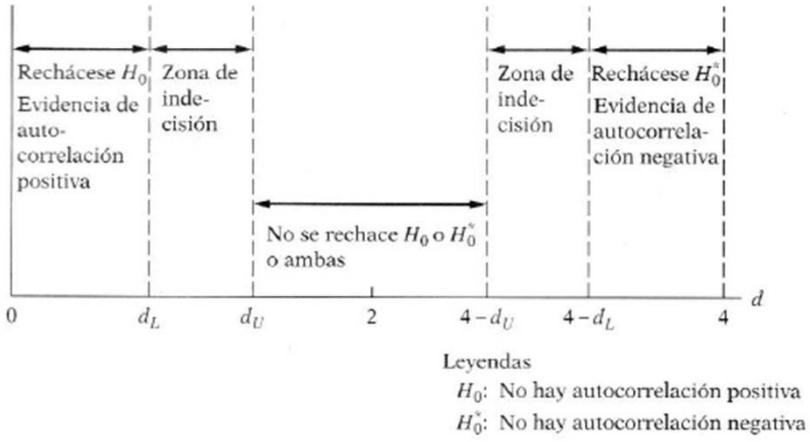
Método Durbin-Watson

Ésta prueba parte de la hipótesis nula de ausencia de autocorrelación de primer orden y define un estadístico de la forma

$$d = \frac{\sum_{t=2}^{n}(\widehat{U}_t - \widehat{U}_{t-1})^2}{\sum_{t=2}^{n}\widehat{U}_t^2}$$

Adicionalmente, Durbin y Watson descubrieron un límite inferior d_L y un límite superior d_u tales que se pudiera concluir acerca de la presencia de autocorrelación serial en el caso en que valor del estadístico cayera por fuera de dichos valores críticos. Estos valores fueron tabulados por los mismos autores en lo que se conoce como las tablas de Durbin-Watson y tienen la ventaja de depender únicamente del número de observaciones y de variables explicativas del modelo.

Para contrastar las hipótesis debe construirse una figura con forma de caja en la que se especifican los límites inferior y superior y las posibles zonas en las cuales puede caer el estadístico. El contraste puede verse ilustrado en la figura:



Así, si el estadístico *d* obtiene un valor cercano a 2, no se presentan problemas de autocorrelación de primer orden. Por el contrario, si obtienen un valor cercano a 0 hay autocorrelación positiva y si presenta un valor cercano a 4 hay autocorrelación negativa.

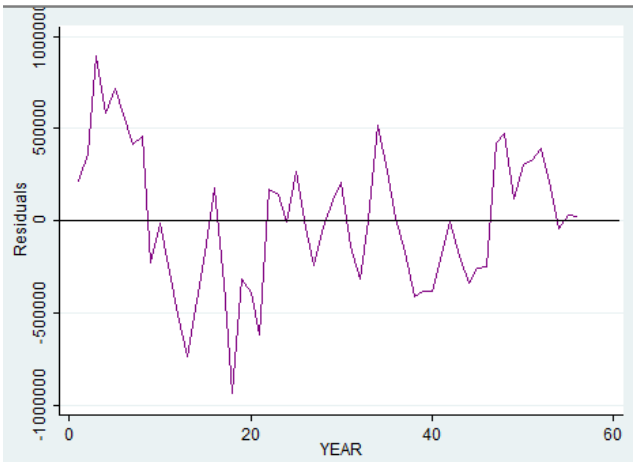
No obstante la prueba detecte exitosamente la presencia de autocorrelación de primer orden, deben tenerse en cuenta una serie de limitantes antes de su aplicación: el modelo a evaluar debe tener término independiente, la prueba no puede aplicarse para detectar autocorrelación de órdenes superiores a 1, existen unas zonas de indecisión en donde no es posible concluir nada acerca de la autocorrelación y por último la prueba no debe aplicarse en modelos que tengan como variable explicativa, la variable endógena rezagada.

2.6.2.1. APLICACIÓN EN SOFTWARE

2.6.2.1.1. Stata 11.0

En Stata, para efectos de que el gráfico de los residuos no presente ningún problema (debido a la periodicidad de los datos), la variable YEAR va a estar conformado por números enteros, viéndose representado cada momento del periodo analizado por un número entre 1 y 56. Es importante recordar que después de efectuar la regresión deben obtenerse los residuos del modelo con el comando **predict** como se explicó en secciones anteriores.

El gráfico se realiza de la misma forma como se realizó el gráfico de dispersión de los residuos en la sección de heteroscedasticidad, a diferencia que en este caso se elegirá una gráfico de línea. Una vez generado el gráfico, para el análisis es necesario añadir una línea correspondiente al valor de 0 de los residuos, con el fin de evaluar la existencia de un comportamiento sistemático. Para esto deberá iniciar el Editor de Gráficos  y allí seleccionar el elemento  , el cual le permitirá agregar una línea en el lugar del gráfico que desee; el resultado será de la siguiente forma:



Estadístico Durbin-Watson

Stata calcula automáticamente el estadístico D-W mediante un comando llamado **estat dwatson**. Sin embargo, para su correcta ejecución es necesario definir la variable YEAR como una variable tipo serie de tiempo, para lo cual se utiliza el comando **tsset** acompañado por el nombre de la variable. A continuación se ilustra el proceso completo:

```
. tsset year
        time variable:  year, 1 to 56
                delta:  1 unit

. estat dwatson

Durbin-Watson d-statistic(  5,    56) =  .6561747
```

Ilustración 60. Estadístico Durbin-Watson-Stata

El resultado arrojado por el método gráfico no es muy preciso ya que no se puede decir objetivamente si se observa o no un comportamiento sistemático; sin embargo el valor del estadístico D-W cae en una zona donde se rechaza la hipótesis nula, por lo que podría concluirse que el modelo presenta autocorrelación positiva de primer orden.

2.6.2.1.2. R-Project

En R-project, el comando utilizado para generar gráficos es **plot()**, el cual en el paréntesis incluye las variables a graficar y una inicial que define el tipo de gráfico que se desee. Para este caso necesitamos generar un gráfico de línea por lo que la inicial será **"l",** la cual se ubicara en el argumento **type**.

Al igual que en Stata, la línea horizontal correspondiente al valor de 0 de los residuos debe añadirse a través de otro comando denominado **abline()**, introduciendo en el paréntesis las coordenadas de ubicación de la nueva línea y su pendiente; adicionalmente el comando ofrece la posibilidad de insertar más fácilmente líneas horizontales o verticales, para las cuales debe escribirse los argumento **h** o **v** seguidos de las coordenadas. Éste comando, como se muestra a continuación, aplicará sus resultados en el gráfico sobre el cual se esté trabajando.

```
> plot(YEAR,res, type="l")
> abline(h=0)
```

Estadístico Durbin-Watson

El comando utilizado para realizar la evaluación del supuesto a través del estadístico D-W, pertenece al igual que uno de los comandos utilizados para evaluar la heteroscedasticidad al paquete **lmtest.** El comando a utilizar en este caso es **dwtest()**, dentro del cual se debe especificar el objeto a evaluar.

```
> dwtest(multiple)

        Durbin-Watson test

data:  multiple
DW = 0.6562, p-value = 1.132e-10
alternative hypothesis: true autocorrelation is greater than 0
```

Ilustración 61. Estadístico Durbin-Watson-R

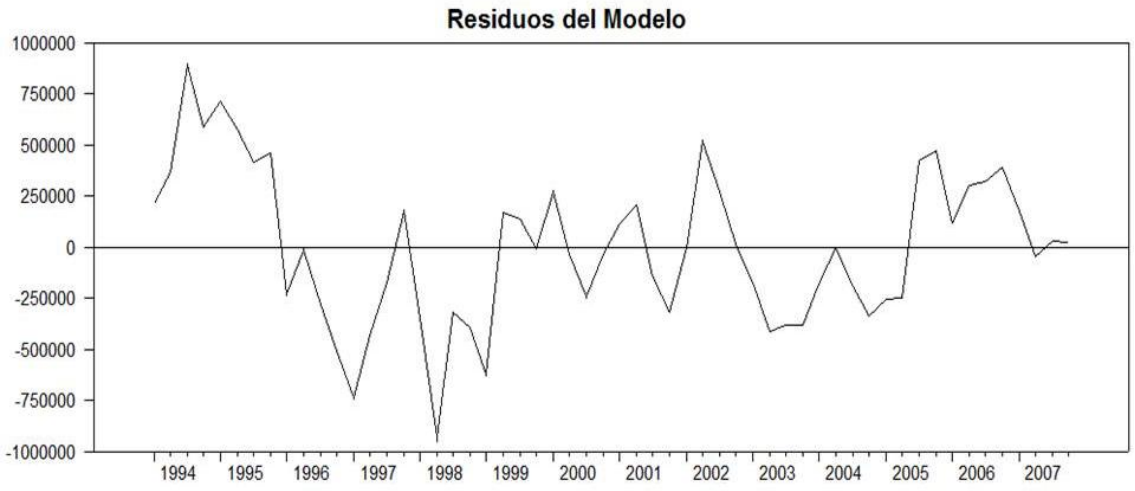
En el resultado se observa el valor del estadístico, su valor-p y la descripción de la hipótesis nula empleada en la prueba.

2.6.2.1.3. WinRATS 7.2

En este software, la aplicación de los métodos para la detección de problemas de autocorrelación es muy sencilla y práctica, puede realizarse rápidamente sin la necesidad de muchos comandos específicos para cada prueba.

En primer lugar, para obtener el gráfico, en la instrucción **graph** debe colocarse únicamente los residuos como variable a graficar, ya que así el software reconoce que la otra variable que la acompaña es el tiempo. Para el ejemplo, se complementó la instrucción con un argumento llamado *header*, el cual se utiliza para colocarle un título al gráfico; el resultado obtenido fue:

```
graph(header="Residuos del Modelo")
# res
```



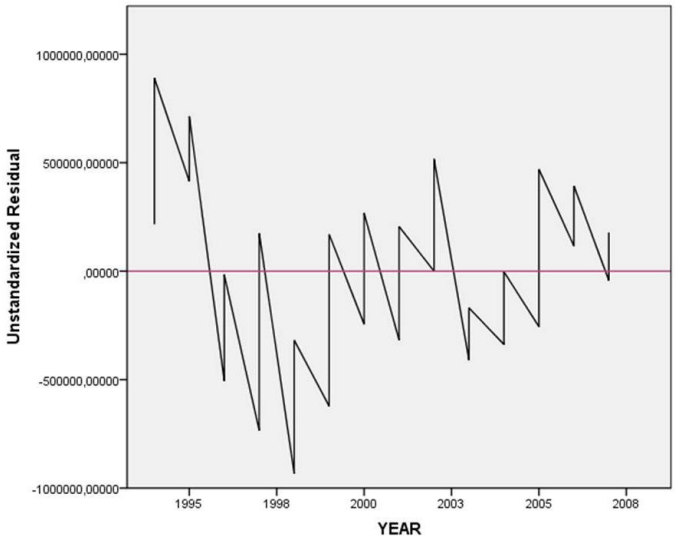
Por otro lado, la obtención del estadístico Durbin-Watson es más sencilla aún, debido a que por defecto, el software ejecuta la prueba al momento de estimar el modelo de regresión lineal, presentando su valor en la parte inferior de la información que este proceso muestra como resultado:

```
Durbin-Watson Statistic            0.656175
```

2.6.2.1.4. SPSS

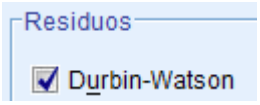
Para emplear el método gráfico en SPSS se debe obtener los residuos del modelo de regresión, recurriendo al mismo proceso explicado en la sección de detección de la heteroscedasticidad. Adicionalmente debe construirse (si no se tiene) en la base de datos una variable que represente los periodos de tiempo.

El gráfico debe construirse con la variable de los residuos en el eje vertical y el tiempo en el eje horizontal. Luego de haberse generado el gráfico en la ventana de resultados, debe agregarse también la línea horizontal en 0, para lo cual el usuario debe acceder al Editor seleccionando el gráfico mediante doble clic. Allí encontrará en la parte superior el siguiente ícono  que le permitirá añadir una línea de referencia al eje Y, para lo cual deberá elegir en la ventana emergente la posición (el valor) sobre el cual desee aplicarla. Siguiendo éste procedimiento, el gráfico resultante se mantiene:



Estadístico Durbin-Watson

En SPSS éste estadístico se considera por defecto como el método básico para la detección de autocorrelación de primer orden, razón por la cual la instrucción para su obtención se hace desde la misma ventana que estima el modelo de regresión. En la opción *Estadísticos*, aparece en la parte inferior, en la sección referente a los residuos del modelo. Para obtenerlo se debe simplemente seleccionar la casilla, como ilustra la imagen, y realizar la regresión.



En la ventana de resultados, el estadístico aparecerá en la última columna de un cuadro llamado *Resumen del modelo*, el cual muestra información adicional sobre todo lo referente al R^2.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,957[a]	,916	,909	387230,5839	,656

a. Variables predictoras: (Constante), IR, PIBR, E, INF
b. Variable dependiente: INVEREX

Ilustración 62. Estadístico Durbin-Watson-SPSS

A través de todos los software se observó que el resultado fue el mismo, por lo que la conclusión acerca de la autocorrelación positiva de primer orden se mantiene.

2.6.3. SUPUESTO DE NORMALIDAD

El supuesto de normalidad en el modelo de regresión lineal implica que los residuos van a estar distribuidos normalmente es decir van a seguir una distribución normal de la forma que su media va a ser cero y su varianza σ_u^2 :

$$U_i \sim N(o, \sigma_u^2)$$

En el caso en que no se cumpla el supuesto de normalidad, los estimadores de MCO no se van a ver afectados, conservando sus propiedades de linealidad, insesgamiento y consistencia. Lo que se va a ver afectado en este caso son las pruebas de inferencia, es decir las pruebas de hipótesis y de significancia debido a que éstas se construyen a partir de la hipótesis de normalidad.

Para contrastar la hipótesis de normalidad pueden utilizarse diferentes herramientas gráficas como lo son el histograma, el diagrama de cajas y el Q-Q plot y otras pruebas más formales como Jarque-Bera y Kolmogorov-Smirnof (prueba no paramétrica no considerada en éste documento).

Jarque-Bera

Se denomina un aprueba asintótica o de grandes muestras lo que quiere decir que a medida que el tamaño de la muestra aumenta, la prueba gana potencia y confiabilidad. Ésta prueba se aplica sobre los residuos contrastando la hipótesis nula de normalidad vs. La alternativa de no normalidad, y parte del cálculo de los coeficientes de asimetría y kurtosis, definidos respectivamente por las ecuaciones:

$$A = \frac{M_3}{M_2^{3/2}} \qquad Ku = \frac{M_4}{M_2^2} \qquad con\ Mj = \sum_{i=1}^{n} \frac{U_i^j}{n} \quad , \qquad j = 1, 2, 3, 4$$

Siguiendo esto, el estadístico Jarque-Bera está definido por la siguiente fórmula

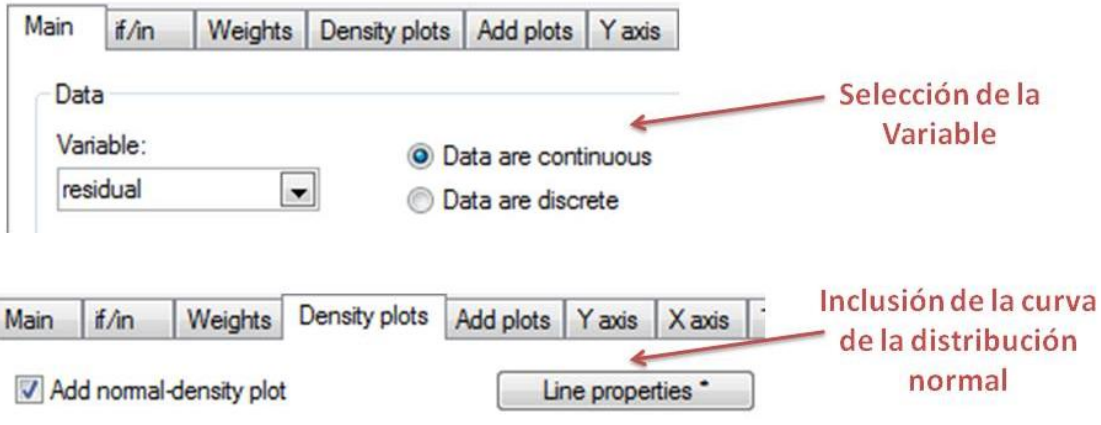
$$J - B = n\left[\frac{A^2}{6} + \frac{(Ku-3)^2}{24}\right]$$

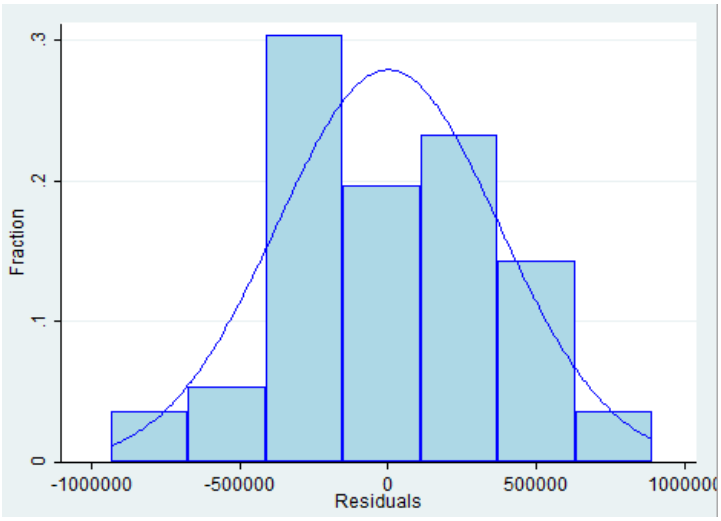
Para tener una distribución perfectamente normal, el coeficiente de asimetría debe ser igual a cero y el de kurtosis igual a 3 por lo que el estadístico J-B tomaría un valor muy cercano o igual a 0. Sin embargo, para contrastar la prueba de hipótesis debe mirarse el valor crítico de la distribución Ji-Cuadrado o el valor-p como se ha hecho anteriormente.

2.6.3.1. APLICACIÓN EN SOFTWARE

2.6.3.1.1. Stata 11.0

Como primera aproximación a la evaluación de la normalidad se realiza el histograma de los residuos. Para esto, se debe recurrir a la pestaña *Graphics > Histogram;* selecciónа partir de la cual aparecerá una ventana en donde principalmente deberá elegirse la variable a graficar y especificar si es continua o discreta. Adicionalmente, Stata ofrece al usuario la facilidad de agregar al gráfico la curva de una distribución normal para comparar directamente si el comportamiento de los residuos se ajusta a ella; operación que se realiza a través de la pestaña llamada *Density plots*. El proceso y resultado se ilustra a continuación en las imágenes





Por otro lado, en Stata se presentan dos alternativas para aplicar la prueba de Jarque-Bera: el test oficial de J-B y otro que al igual que éste se basa en el cálculo de la asimetría y kurtosis.

Para la primera prueba, es necesario instalar un paquete llamado **JB6** lo cual puede hacerse a través del comando **ssc install** como se vio en la sección dedicada al análisis del cambio estructural. El comando para aplicar la prueba es **jb** y va acompañado del objeto a evaluar,

```
. jb residual
Jarque-Bera normality test:   .1283 Chi(2)   .9379
Jarque-Bera test for Ho: normality:
```

Ilustración 63. Jarque-Bera.Stata

La segunda prueba si se encuentra dentro de las pruebas predeterminadas del software y se utiliza igual a la anterior, escribiendo el código y a continuación la variable. La prueba recibe el nombre de **sktest** y presenta por separado los resultados de un test de normalidad basado en la kurtosis, otro basado en la asimetría y la combinación de los dos, como se ilustra en la figura

```
. sktest residual
```

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
residual	56	0.9469	0.9731	0.01	0.9972

Skewness/Kurtosis tests for Normality

Ilustración 64. Test alternativo de normalidad-Stata

Ambas pruebas muestran un valor-p que es mucho mayor al valor de α, por lo que se puede concluir que para el modelo que se está analizando los errores se distribuyen normalmente.

2.6.3.1.2. R-Project

En R existe un comando especialmente diseñado para obtener histogramas, el cual es muy útil para la evaluación de la normalidad. Así, para obtener el histograma de los residuos, se utiliza el comando **hist()**, colocando en el paréntesis la variable a graficar, como se ilustra.

```
> hist(res)
```

Por otro lado, para aplicar la prueba J-B en R se necesita la previa instalación del paquete **tseries**. El comando a utilizar es muy intuitivo ya que se reconoce por el mismo nombre de la prueba: **jarque.bera.test (x),** donde x es el objeto a analizar, es decir los residuos. Para el caso ilustrativo, el comando fue aplicado de la siguiente manera,

```
> jarque.bera.test(res)

        Jarque Bera Test

data:  res
X-squared = 0.1283, df = 2, p-value = 0.9379
```

Ilustración 65. Jarque-Bera-R

Mostrando como resultado el valor del estadístico, los grados de libertad utilizados y el valor-p para contrastar la prueba de hipótesis.

2.6.3.1.3. WinRATS 7.2

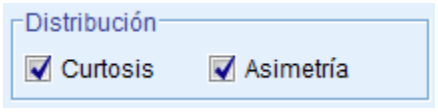
La evaluación del supuesto de normalidad en RATS es muy sencillo, ya que la obtención del estadístico J-B no necesita de ningún comando en específico o de la construcción de una fórmula que permita calcularlo, por el contrario este se encuentra por defecto en las estadísticas básicas de los residuos del modelo. Para visualizarlo, se necesita entonces obtener información sobre los residuos, proceso que, el usuario recordará, se lleva a cabo por medio del comando **statistics**. Debido a que solo se desea conocer información sobre un componente del modelo, el comando fue utilizado de la siguiente manera:

```
statistics res

Jarque-Bera   0.048186      Signif Level (JB=0) 0.976195
```

2.6.3.1.4. SPSS

En este software la prueba debe ser construida siguiendo la fórmula con la cual está definido el estadístico. Como primera medida debe obtenerse información acerca de la asimetría y kurtosis de los residuos, para lo cual debe dirigirse a la pestaña *Analizar* > *Estadísticos Descriptivos* > *Descriptivos*, allí seleccionar la variable de los residuos y finalmente seleccionar el botón "Opciones", en donde podrá seleccionar la información mencionada anteriormente, así:



Una vez obtenida esta información (encontrada en la ventana de resultados), puede procederse a calcular el estadístico a través de la pestaña *Transformar* > *Calcular variable…*. La fórmula construida a modo de ilustración fue la siguiente:

JB = 56 * (((0.02026754863021407) ** 2 / 6) + ((-0.13786938871707194) ** 2 / 24))

donde el valor 0.02026 corresponde a la asimetría y el valor -0.13786 a la kurtosis. De esta forma, el valor calculado del estadístico fue 0,0481. A continuación se obtuvo el valor-p del estadístico Ji-Cuadrado construyendo la fórmula

1 - CDF.CHISQ(JB,2)

El valor obtenido fue 0,9761.

Aunque los resultados de las pruebas no fueron exactamente los mismos en todos los software, la conclusión acerca de que los residuos siguen una distribución normal no se ve alterada.

2.6.4. COMPARACIÓN DE LAS PRUEBAS DE HIPÓTESIS SOBRE LOS ERRORES EN CADA SOFTWARE

Al igual que en la sección dedicada a contrastar los supuestos sobre la estructura del modelo, a partir de la aplicación de pruebas para validar los supuestos acerca de la perturbación aleatoria del modelo, es posible hacer una tabla comparativa que resalte las características de cada software.

SOFTWARE	HOMOSCEDASTICIDAD	NO AUTOCORRELACIÓN	NORMALIDAD
Stata	Comandos precisos y fáciles de identificar para la aplicación de todas las pruebas.	Contraste preciso. Necesidad de paquete adicional para su aplicación.	Variedad de pruebas de sencilla aplicación para el contraste del supuesto.
R-Project	Diferentes alternativas para la aplicación de una misma prueba. Confusión en los códigos que las identifican.	Contraste preciso. Necesidad de paquete adicional para su aplicación.	Contraste preciso. Necesidad de paquete adicional para su aplicación.
WinRATS	Construcción simple de los estadísticos de las pruebas y su contraste.	Generación automática del estadístico necesario sin necesidad de códigos adicionales al proceso de generación de la regresión.	Obtención sencilla de la prueba sin la necesidad de su construcción.
SPSS	Necesidad de una construcción elaborada y compleja de los	Aplicación sencilla de la prueba. Fácil obtención de	Necesidad de construcción de la prueba a partir de

	estadísticos empleados por las pruebas debido a la inexistencia de opciones para la evaluación del supuesto.	resultados.	información provista por las estadísticas.

Tabla II. Comparación de las pruebas de hipótesis sobre los errores en cada software

3. CONCLUSIONES

A lo largo del documento se hizo por un lado, una revisión teórica acerca del modelo de regresión lineal, sus supuestos, implicaciones e importancia en su aplicación para la simplificación y entendimiento de los variados fenómenos que se presentan en la realidad. Por otro lado se hizo una revisión práctica de las herramientas ofrecidas por 4 de los software más reconocidos a nivel mundial en el área de la estadística y la econometría: Stata, R-Project, winRATS y SPSS, con el fin de evaluar las características, ventajas y desventajas de cada uno de ellos.

Conocer la teoría detrás del método de regresión lineal es un asunto fundamental para todos los estudiantes de economía como parte de su formación profesional, sin embargo no puede dejarse a un lado la aplicación práctica de todo el conocimiento y las herramientas aprendidas en clase para la resolución de los problemas a los cuales se van a ver enfrentados en el futuro. Adicionalmente, en respuesta a la rápida evolución de la tecnología, es imprescindible que tanto estudiantes como docentes se encuentren a la vanguardia de las herramientas informáticas que están a su disposición con el objetivo de destacarse en el mundo académico y laboral.

Por estas razones, se realizó la exploración de todo lo referente al método de regresión lineal en los software anteriormente mencionados, permitiendo así proporcionar a la comunidad de la Facultad de Ciencias Económicas y especialmente a los Economistas un soporte y acompañamiento durante el desarrollo de su carrera. En el documento encontrarán una serie de instrucciones que les permitirán desenvolverse con facilidad en la utilización de cada programa, siendo esto un complemento a su formación y una ampliación de su perspectiva en lo referente a las tecnologías que son adoptadas diariamente en el mundo real.

En cuanto a la evaluación de cada uno de los software por separado, se observó que si bien todos responden satisfactoriamente al objetivo de la aplicación del método de regresión

lineal, cada uno presenta una serie de ventajas y desventajas en cuanto a su utilización y las herramientas que ofrece, cuestión que puede ser decisiva para el usuario al momento de elegir el software de su predilección. Las principales ventajas y desventajas encontradas por el autor se ven representadas en la Tabla I.

SOFTWARE	VENTAJAS	DESVENTAJAS
Stata	• Interfaz amable al usuario. • Trabaja por medio de menús desplegables y un sencillo lenguaje de programación. • Extensibilidad a través de paquetes • Muy buen contenido gráfico • Compatibilidad con todos los sistemas operativos.	• Software privativo • Limitaciones para la instalación de paquetes.*
R-Project	• Software libre y altamente extensible • Muy buen contenido gráfico. • Compatibilidad con todos los sistemas operativos. • Gran cantidad de documentación.	• Código que para ciertos usuarios puede ser considerado complejo.
WinRATS	• Requiere la programación de todas las pruebas, lo que permite al usuario entenderlas mejor. • Compatibilidad con todos los sistemas operativos.	• Software privativo • Falta de claridad en el contenido de ayuda.
SPSS	• Interfaz intuitiva y amable al usuario. • Trabaja a partir de menús desplegables. • Fortaleza en análisis de datos y análisis estadístico. • Compatibilidad con archivos de otras extensiones.	• Software privativo • Falta de herramientas para un análisis econométrico completo.

*Aplica únicamente para las instalaciones de la Universidad Nacional de Colombia debido a cuestiones del proxy.

Tabla III. Ventajas y Desventajas de los Software

Así, el presente documento presenta a los lectores varias herramientas informáticas para la resolución de un mismo problema, con el objetivo de servir como complemento a la formación integral de los estudiantes de la facultad, ampliando su visión acerca de la gran variedad de instrumentos a su disposición.

4. REFERENCIAS

- Estadística Descriptiva. Revisado el 26 de Agosto de 2011. Dirección URL: http://sitios.ingenieria-usac.edu.gt/estadistica/estadistica2/estadisticadescriptiva.html
- Gujarati, Damodar. (2003). *Econometría.* Editorial McGraw-hill-México, 4ª edición.
- Apuntes de clase *Econometría I.* Profesor: Hector Cárdenas. Universidad Nacional de Colombia.
- Medidas Descriptivas. Revisado el 26 de Agosto de 2011. Dirección URL: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema3.pdf
- Revisado el 26 de Agosto de 2011. Dirección URL: http://www.tuveras.com/estadistica/estadistica02.htm
- Revisado el 26 de Agosto de 2011. Dirección URL: http://www.ematematicas.net/estadistica/medidas/index.php
- Revisado el 26 de Agosto de 2011. Dirección URL: http://www.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/18reglin_SPSS.pdf
- R-Project. Revisado el 27 de Agosto de 2011. Dirección URL: http://www.r-project.org/
- STATA, data analysis and statistical software. Revisado el 26 de Agosto de 2011. Dirección URL; http://www.stata.com/whystata/
- StataCorp (2007). *Getting started with Stata for Windows, release 10.* Stata Press.
- IBM SPSS Statistics: Capabilities. Revisado el 10 de Septiembre de 2011. http://www-01.ibm.com/software/analytics/spss/products/statistics/capabilities.html
- SPSS Tutorial. http://127.0.0.1:4627/help/index.jsp?topic=/com.ibm.spss.statistics.tut/introtut2.htm.
- División económica. Departamento de investigaciones económicas. *Principales indicadores para el diagnóstico del análisis de regresión lineal.* Revisado el 30 de Septiembre de 2011. http://www.bccr.fi.cr/ndie/Documentos/DIE-37-2003-IT-INDICADORES%20PARA%20ANALISIS%20DE%20REGRESION.pdf

- Perez, Blanca & García, Maria. (2010). *Análisis del cambio estructural en el Modelo de regresión lineal.* Revisado el 15 de Noviembre de 2011. http://www.latindex.ucr.ac.cr/mate-17-2/matematica-17-2-06.pdf
- Estima. RATS. http://www.estima.com/ratsmain.shtml
- OCW Universidad de Cantabria (2008). *Capítulo 5.* http://ocw.unican.es/ciencias-sociales-y-juridicas/econometria/econometria/apuntes/tema5.pdf
- Melo, Luis & Misas, Martha. *Modelos Estructurales de Inflación en Colombia: Estimación a través de Mínimos Cuadrados Flexibles* http://banrep.org/docum/ftp/borra283.pdf
- Stata. *How Can I Compute the Chow test Statistic?* http://www.stata.com/support/faqs/stat/chow.html
- Estima.(2007) *RATS Version 7, Reference Manual.* http://digidownload.libero.it/rocco.mosconi/Ref_man_RATS.pdf.
- *CUSUM tests*. http://personal.rhul.ac.uk/uhte/006/ec5040/Cusum%20test.pdf
- Acuña, Edgar. *Multicolinealidad*. math.uprm.edu/~edgar/cap7sl.ppt
- UCLA Academic Technology Services. *Stata Web Books, Regression with Stata, Chapter 2-Regression Diagnostics.* http://128.97.141.26/stat/stata/webbooks/reg/chapter2/statareg2.htm
- Stata 9.2. *Comandos Útiles.* http://www.ugr.es/~montero/matematicas/stata.pdf
- *Multicolinealidad*.http://www.uv.es/uriel/material/multicolinealidad3.pdf
- *Multicolinealidad en el MLG.* http://dae.unizar.es/monia/tema%202_%20MULTICOLINEALIDAD%20EN%20EL%20MLG.pdf
- http://www.bsos.umd.edu/gvpt/glayman/heteroskedasticity_examples.pdf
- *Kellog School of Management.* Introduction to SPSS. www.kellogg.northwestern.edu/kis/tek/ongoing/Materials/Introduction2SPSS.pdf

5. INFORME DE ACTIVIDADES

(Este apartado del documento debe ser eliminado cuando se apruebe la publicación del documento, entre tanto, su finalidad es la de dar seguimiento al proceso de investigación)

Actividad	% de Cumplimiento
Análisis descriptivo de los datos en los software	100 %
Avance de los métodos de ajuste y estimación de modelos	100 %
Revisión de alternativas para la evaluación de hipótesis de Muestras Pequeñas	100 %
Revisión de alternativas para la evaluación de hipótesis de Cambio Estructural	100%
Revisión de alternativas para la evaluación de hipótesis de Especificación Erronea	100%
Revisión de alternativas para la evaluación de hipótesis de Multicolinealidad	100%
Revisión de alternativas para la evaluación de hipótesis de HOMOCEDASTICIDAD	100%
Revisión de alternativas para la evaluación de hipótesis de NO AUTOCORRELACIÓN	100%
Revisión de alternativas para la evaluación de hipótesis de NORMALIDAD	100%
Desarrollo de la aplicación teórico-práctica con miras a ser publicada en el Journal UIFCE	
Total	80 %