MANUAL: Econometría aplicada utilizando R

Madin Rivera, Alberto.

may. 07, 2023

Índice general

Este manual es una readaptación del texto original del libro creado por los Coordinadores[1]:

- Luis Quintana Romero
- Miguel Ángel Mendoza González

Docéntes a los cuáles también se les debe de reconocer y dar el agradecimiento total son:

- Javier Galán Figueroa
- Jorgue Feregringo Feregringo
- Lucía A. Ruíz Galindo
- Roldan Andrés Rosales

Esto es debido a que el crédito original, del cuál se readaptó este manual es gracias a todos ellos.

[1]Este manual tiene Prohibida la reproducción total o parcial por cualquier medio sin la autorización escrita del titular de los derechos patrimoniales. El libro electrónico *Econometría aplicada utilizando R* fue financiado con recursos PAPIME de la Dirección General de Asuntos del Personal Académico (DGAPAPA) de la Universidad Nacional Autónoma De México: PE302513 *Libro electrónico y complementos didácticos en medios computacionales, para el fortalecimiento en la enseñanza de la econometría.* Se encuentra disponible de manera libre en el sitio: Econometría aplicada Utilizando `R`

INTRODUCCIÓN

En este libro de texto los usuarios encontrarán una vía práctica para mejorar su comprensión de la econometría, al utilizar aplicaciones a su realidad social, emplear las fuentes de información disponibles en el país y disponer de un formato tecnológico en el que pueden aplicar los conocimientos adquiridos, poner en práctica propuestas propias y realizar trabajos de investigación por su cuenta, haciendo uso de medios tecnológicos de uso masivo.

Los capítulos de este libro de texto tienen como eje común la aceptaci´ón de que en los últimos veinte años se ha dado una revolución en las técnicas econométricas y en sus aplicaciones. Es buena parte estos cambios, provienen del reconocimiento de que el paradigma clásico, que actualmente aún predomina en la mayoría de los libros de texto, fue sustentado en supuestos muy discutiples. Los cuestionamientos a la metodología econométrica clásica se desprenden del trabajo de Box y Jenkins (1970) en **series de tiempo**; Davidson, Hendrym Srba y Yeo (1978) que desarrollaron la idea de **modelos de corrección de error** (MCE) y que actualmente su propuesta se reconoce como metodología LSE (London School of Economics) o DHDY (por las iniciales de sus autores); los numerosos trabajos de Engle y Granger a partir de los años ochenta en donde se vincula el concepto de cointegración a los MCE; el trabajo del mismo Engle (1982) que dio lugar a los modelos ARCH (heterocedasticidad condicional autorregresiva), los cuáles han tenido un gran impacto en el análisis econométrico aplicado al mundo de las finanzas; los desarrollos de finales de los años noventa en el campo de la Econometría Espacial impulsados por Anselin (1988) y; un sin número de artículos que inspirados en estos trabajos pioneros han cambiado la forma de pensar y hacer econometría en la actualidad.

El reto de este libro es ofrecer a los lectores un enfoque aplicado con el fin de comprender esos nuevos desarrollos en el campo de la econometría y proporcionarles las herramientas teóricas y las técnicas necesarias para su aplicación al estudio de la realidad económica mexicana.

Los libros de texto de econometría que se están publicando recientemente, tanto en Europa como en los Estados Unidos, se vinculas a paquetes computacionales de elevado costo comercial, como lo son: `EViews`, `STATA` y `Microfit`, entre otros. Sin embargo, actualmente se ha desarrollado software de uso libre que ha adquirido una gran difusión mundial (como `Python` o `R`). `R` es uno de ellos, el cual se ha venido utilizando para la modelación econométrica con mucho éxito.

Por tal razón, el presente libro de texto de econometría tiene la peculiaridad de que utiliza ampliamente los desarrollos disponibles libremente en `R`, además de priorizar la aplicación de los temas que se desarrollan en sus diferentes capítulos. En cada uno de los capítulos del libro se muestran las bases del método o técnica econométrica de que se trate y se aplica inmediatamente al estudio de algún tema relevante de la economía mexicana actual o de otros países.

Los capítulos que conforman este libro presentan un nivel introductorio de cada uno de los temas que se abordan y se priorizan las aplicaciones en `R`, por lo cual debe considerarse como un libro de econometría básica aplicada. Se ha dejado fuera del texto el tema de los modelos de **series de tiempo**, ya que por la amplitud de ese tema se requiere de un libro adicional, mismo queya se encuentra en proceso de preparación con el fin de complementar a la presente obra.

Se debe de señalar que este libro de texto forma parte de la producción y edición de tres

materiales educativos en el campo de la econometría. Los materiales consisten de un libro electrónico (ebook) de texto, un curso en línea y aplicaciones electrónicas didácticas.

Estos materiales están destinados a profesores y alumnos. En el caso de los profesores es pisible emplear el texto electrónico y el curso en línea para cursos de actualizaci+on del personal docente en econometría. Los profesores pueden utilizar los materiales en la impartición de cursos a nivel licenciatura, ya que los materiales se diseñan de acuerdo a los contenidos de los programas curriculares de econometría y de métodos de pronóstico en diferentes licenciaturas, resolviendo con ello el déficit existente de material actualizado, en español, en soportes electrónicos y con aplicaciones a la realidad del país.

La propuesta es original en la medida en que atiende tres problemas de la enseñanza de la econometría; contar con libros de texto actualizados en formatos tecnológicamente avanzados y en español, incorporar un curso en línea que tenga la virtud de promover el auto aprendizaje y sea complemento de los cursos presenciales, además de proporcionar aplicaciones en formatos tecnológicos que se han difundido ampliamente entre los alumnos.

Los materiales vinculados a este libro de texto se encuentra disponible de forma libre en la página de la UNAM. En este sitio, el interesado en el estudio de la econometría, encontrará este libro en formato electrónico, presentaciones en PowerPoint para cada capítulo, una grabación de video con los procedimientos para aplicar en `R` lo aprendido en el capítulo, una guía metodológica en MOODLE para avanzar en el estudio de los capítulos y, finalmente, un par de aplicaciones electrónicas para comprender la forma en la que se estiman regresiónes.

El libro se integra por dieciséis capítulos cuyo contenido se resumen en la siguiente tabla.

CAPÍTULO	CONTENIDO
CAPÍTULO 1.	Se introduce al lector en la metodología econométrica moderna y en el uso del R
CAPÍTULO 2.	Se muestra el método de mínimos cuadrados ordinarios en su versión matricial con ejemplos de análisis de la deuda pública en México
CAPÍTULO 3.	Se desarrolla el modelo de regresión múltiple y la forma en la cual se evalúan sus resultados. Se realizan aplicaciones en R al análisis de las ventas al menudeo en México
CAPÍTULO 4.	Se presentan los métodos utilizados para determinar si el modelo econométrico fue especificado incorrectamente debido a un planteamiento no apropiado de la forma funcional. Se realizan aplicaciones en R con el análisis de la demanda de gasolina en los Estados Unidos
CAPÍTULO 5.	En este capítulo se estudia la importancia e implicaciones del supuesto de normalidad en el modelo de regresión lineal y de manera específica en la inferencia estadística de sus parámetros. Se realizan aplicaciones en R de la prueba Jarque-Bera en un modelo de la demanda de gasolina en los Estados Unidos
CAPÍTULO 6.	Con base en los determinantes del consumo en México se exploran las diferentes pruebas alternativas disponibles en R para detectar y corregir el problema de la multicolinealidad en los modelos econométricos
CAPÍTULO 7.	Se explican las consecuencias del problema de heterocedasticidad en los modelos econométricos y haciendo uso de un ejemplo sobre distribución de cerveza se muestran las alternativas disponibles en R para realizar pruebas de detección de ese problema
CAPÍTULO 8.	La autocorrelación serial y sus consecuencias es analizada con base en el estudio de las tasas de interés en México. Utilizando R se muestran las pruebas para detectar este problema y las alternativas para su solución

CAPÍTULO	CONTENIDO
CAPÍTULO 9.	En este capítulo se aborda uno de los temas más relevantes de la metodología econométrica moderna que es el de identificar el orden de integración de las variables utilizadas en los modelos econométricos. Con base en el R se realizan pruebas de raíz unitaria utilizando como ejemplo el análisis del Producto Interno Bruto de México
CAPÍTULO 10.	Los resultados del capítulo anterior se extienden al estudio de los procesos de cointegración entre las variables del modelo econométrico utilizando en R las técnicas de Engle-Granger y de Johansen, ejemplificándolas con ayuda del estudio de la relación de largo plazo entre el consumo y el ingreso en México
CAPÍTULO 11.	Se destaca el uso de modelos VAR para el análisis de la política económica tomando como caso el estudio de la inflación y la oferta monetaria. Se presentan las diferentes rutinas disponibles en R para estimar y realizar pruebas en los modelos VAR
CAPÍTULO 12.	Los modelos ARCH utilizados para el análisis de la volatilidad y el riesgo son ejemplificados en R con base en el análisis de los procesos inflacionarios en México
CAPÍTULO 13.	Se desarrollan los modelos Probit y Logit aplicados a casos en los que la variable dependiente es binaria o cualitativa. Con base en el estudio de la diferenciación salarial en México se muestran las rutinas disponibles en R para estimar y realizar pruebas en ese tipo de modelos econométricos
CAPÍTULO 14.	Cuando el fenómeno económico. que se está analizando tiene un componente de desagregación de corte trasversal o sección cruzada y otro de series de tiempo se aplican modelos de panel. En este capítulo se estudian las técnicas de panel utilizando R en el análisis de la inflación y el desempleo en México

CAPÍTULO	CONTENIDO
CAPÍTULO 15.	Uno de los desarrollos más recientes de la econometría es la econometría espacial. En este capítulo se presenta la forma en la que se deben especificar y estimar este tipo de modelos en R y se ejemplifica su uso con el estudio del empleo y el capital humano en la zona centro de México
CAPÍTULO 16.	Finalmente, se incluye un capítulo opcional en el que se realiza un breve repaso de los elementos básicos de estadística, probabilidad y algebra lineal indispensables para comprender la base matemática de los diferentes capítulos del libro

Este libro y los materiales didácticos adicionales que lo acompañan contaron con el apoyo financiero de la Dirección General de Asuntos del Personal Académico de la UNAM a través del proyecto PAPIME PE302513 *"Libro electrónico y complementos didácticos en medios computacionales, para el fortalecimiento en la enseñanza de la econometría"*.

Los coordinadores del libro agradecen a los profesores José A. Huitrón, Jaime Prudencio, Aída Villalobos y Ángel Reynoso por su apoyo en la revisión de los capítulos y en el diseño de los apoyos didácticos que acompañan al libro. También agradecemos a los alumnos y becarios del proyecto PAPIME; Arturo Abraham Salas, Mónica González, Paola Orozco, Ana Isabel Hernández, Coral Gutiérrez, Eddy Michell López, Jarett Fernando González, Mónica Patricia Hernández, Samarkanda Norma Bustamante, Nataly Hernández, Sarahí Aldana, Brenda Mireya González, Alejandro Corzo, Damaris Susana Mendoza, Nancy Nayeli Morales, Claudia Torres, Edelmar Morales y Carolina Guadalupe Victoria. Todas y todos ellos hicieron una excelente labor de apoyo para el buen éxito del proyecto.

CAPÍTULO 1: LA ECONOMETRÍA. SUS USOS Y APLICACIONES EN R

- Por Quintana Romero, Luis. y Mendoza, Miguel Ángel.

1. ¿QUÉ ES LA ECONOMETRÍA?

Hoy en día la econometría se ha difundido ampliamente entre quienes estudian y buscan realizar aplicaciones de la economía. En general, cualquier licenciatura en economía cuenta, entre su currículo, con uno o más cursos de econometría; hoy en día es usual que la econometría se enseñe con la misma relevancia que se le da a los cursos de microeconomía y macroeconomía. No hay posgrado en economía que deje de incorporar el estudio de la econometría como una disciplina fundamental. Incluso, es posible aseverar que en disciplinas distintas a la economía, como en las matemáticas, algunas ingenierías, la sociología y en la psicología, sus estudiantes reciben algún curso de econometría.

No sólo en la formación académica la econometría está presente, en la vida laboral se realizan todos los días aplicaciones econométricas. En las oficinas gubernamentales se emplean modelos econométricos para realizar pronósticos de variables económicas. En empresas privadas se utilizan algunas técnicas econométricas para proyectar al futuro variables como ventas, precios y demanda, entre otras variables. En el mercado existen numerosos servicios de consultoría que han hecho de la econometría un negocio al ofrecer la venta de pronósticos generados a través de modelos econométricos.

En el mundo de la investigación científica la econometría es un ingrediente indispensable. Diariamente se publican en todo el orbe una gran cantidad de artículos de economía en revistas especializadas, la evidencia empírica que aportan, generalmente, se sustenta en algún modelo econométrico.

La importancia de esta disciplina es tal que basta escribir en un buscador de internet la palabra *"econometrics"*, para que nos arroje más de nueve millones de referencias.

Con la econometría se busca comprender fenómenos como el de las crisis, identificar sus causas, valorar sus consecuencias futuras y proponer medidas de política para enfrentarlas. Para ello, la econometría utiliza modelos, con estos se busca representar de forma simplificada a los principales factores causales de un problema de interés. La especificación y estimación de esos modelos requiere del conocimiento de teorías económicas, para poder establecer relaciones entre las variables, y de datos, para poder realizar mediciones de dichas relaciones.

No existe una definición única y generalmente aceptable de lo qué es la econometría. Debido a que en ella concurren una gran diversidad de perspectivas teóricas y metodológicas, existen, en consecuencia, diferentes posturas sobre su significado.

A diferencia de lo que ocurre hoy en día, en los años treinta, época en la que se institucionaliza la econometría, existía cierto consenso metodológico. A ese consenso se le identifica como la *"metodología de libro de texto"* y su definición de econometría era la siguiente:

> *"La aplicación de métodos estadísticos y matemáticos al análisis de los datos económicos, con el propósito de dar un contenido empírico a las teorías económicas y verificarlas o refutarlas" (Maddala, 1996, p.1)*

Bajo esta última conceptualización la econometría aparece, por un lado, como un mero instrumental técnico al ser la aplicación de métodos matemáticos y estadísticos. Por otro lado,

es vista prácticamente como la piedra filosofal, al darle el papel de criterio último de verdad al ser la vía para verificar o refutar teorías. El econometrista aparece en esa definición como un técnico, cuyo único fin es intentar medir lo que la teoría económica ha postulado.

Esta visión de la econometría se ha transformado en los últimos años, en ese sentido vale la pena retomar la definición proporcionada por Aris Spanos:

> *"La econometría se interesa por el estudio sistemático de fenómenos económicos utilizando datos observables" (Spanos, 1996, p.3).*

Este es un enfoque moderno con el cual se coincide en este libro, lo que hace a la econometría diferente de otros campos de la economía es la utilización de datos observables. Por lo tanto, la econometría tiene una perspectiva empírica, no se reduce a la teoría y necesariamente hace uso de datos, los cuales no son experimentales sino que son resultado del funcionamiento de la actividad económica. El papel del econometrista no se reduce a medir lo que la teoría económica establece, es un científico social que, a través de un método científico, emprende el estudio de fenómenos económicos. Por lo tanto, no es un observador pasivo de la teoría, al contrario, es capaz de contribuir a la teoría.

La econometría que utilizamos hoy en día se ha ido transformando y modernizando, hasta convertirse en una de las herramientas más potentes a disposición de los economistas y principalmente del análisis empírico de problemas económicos. Esta evolución de la disciplina la sintetiza perfectamente Spanos:

> *"En el amanecer del siglo veintiuno, la econometría se ha desarrollado desde los modestos orígenes del"ajuste de curvas" por mínimos cuadrados en los inicios del siglo veinte, hasta un poderoso arreglo de herramientas estadísticas para modelar todo tipo de datos, desde las tradicionales series de tiempo a las secciones cruzadas y los datos de panel." (Spanos, 2006, p. 5)*

2. LA METODOLOGÍA ECONOMÉTRICA

En el apartado previo se estableció que la econometría estudia de forma sistemática los fenómenos económicos. Por lo tanto, utiliza una metodología científica para llevar a cabo esta tarea. Aunque la metodología econométrica no tiene aún un lugar relevante en la discusión de esta disciplina, es un aspecto que debe ser considerado esencial, por ello resulta muy atinada la afirmación de Spanos (2006) en el sentido de que sin fundamentos metodológicos para guiar la práctica econométrica, no es posible que se logre acumular conocimiento genuino a través de la modelación empírica.

En la medida en que existe una diversidad metodológica en la econometría, resulta difícil establecer un proceso metodológico único. Sin embargo, en términos generales, en el cuadro siguiente se pueden observar las características básicas de los principales enfoques metodológicos, los cuales se distinguen por el papel que le asignan a la teoría y del grado de independencia que le dan a la teoría para la caracterización de los datos Hoover (2006)[2].

[2]Fuente: Elaboración propia con base en Hoover (2006)

TABLA 1: PERSPECTIVAS METODOLÓGICAS EN LA ECONOMETRÍA

METODOLOGÍA	PERIODO	AUTORES	CARACTERÍSTICAS
Comisión Cowles	Años 40 y 50	Koopmans	Se centró en el problema de identificación y el papel de la teoría para establecer las restricciones de identificación
Vectores Auto Regresivos (VAR)	Años 80	Christoper Sims	Enfoque sin tería en la estructura de los datos y uso de ecuaciones VAR para modelar impactos en las variables
Calibración	Años 90	Finn Kydland y Edward Prescott	Modelos teóricos de expectativas racionales a los que se les asignan valores numéricos en los parámetros claves
Libro de texto	Años 90 y 2000	Post Comisión Cowles	Resurge la metodología de la Comisión Cowles aplicada a modelos uniecuacionales con métodos instrumentales
London School Economics (LSE)	Años 90 y 2000	Denis Sagan, David Hendry	Especificaciones dinámicas, cointegración y búsqueda de especificaciones parsimoniosas; Anidamiento y metodología de lo general a lo específico

Dentro de estas perspectivas la LSE ha jugado un papel destacado al contraponerse a la de libro de texto y conformar lo que puede denominarse una nueva metodología econométrica. La de libro de texto parte del supuesto de que el modelo teórico es el verdadero modelo y, en consecuencia, coincide con el proceso generador de los datos (PGD). En consecuencia, para esa metodología, a econometría se reduce a la estimación de los parámetros que la teoría plantea; mide lo que la teoría dice, pero no explica nada.

Al contrario, la LSE parte de la idea de que los modelos son aproximaciones teóricas y empíricas del PGD. La validación de esas aproximaciones se realiza a través de la evaluación de los modelos utilizando una amplia batería de pruebas estadísticas que buscan determinar la congruencia de esas aproximaciones con el PGD. El PGD como fenómeno económico de interés que da lugar a los datos, no es conocido debido a que los datos son observacionales y no experimentales; los datos que se utilizan en los modelos econométricos no son generados en un laboratorio bajo control[3].

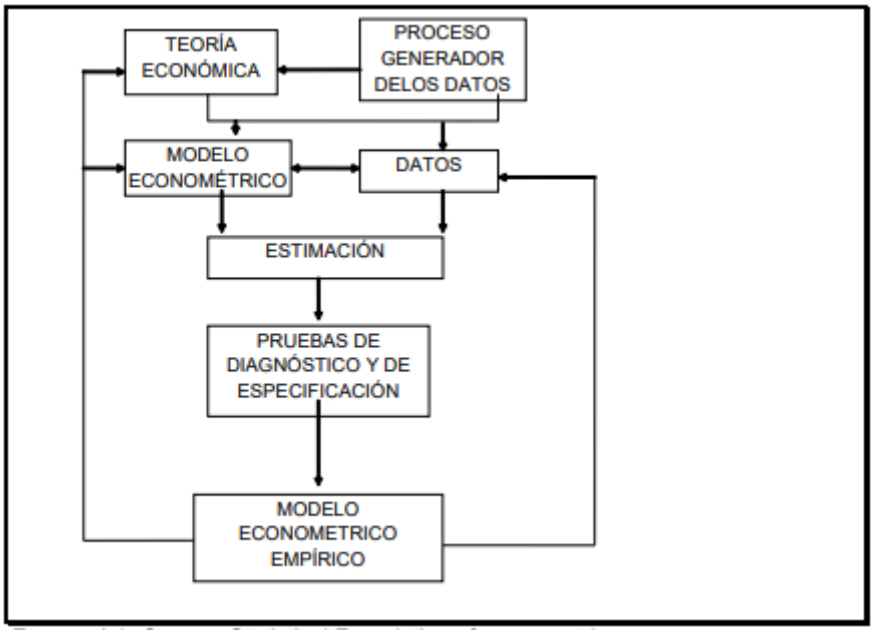


Figura 1: Nueva Metodología econométrica

[3]Fuente: Aris Spanos *Statistical Foundation of econometrics*

3. EL MODELO ECONOMÉTRICO

Los modelos econométricos son una simplificación de la realidad que se compone de relaciones entre variables. Dichas relaciones son no exactas y, por ello, se les llama relaciones estadísticas y pueden describirse en términos probabilísticos. Este tipo de relaciones funcionales pueden expresarse como un modelo estadístico para una variable dependiente y_i y un conjunto de $k-1$ variables explicativas o regresores X_{ki}:

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \tag{1}$$

En donde el término u_i es un error o pertudbación aleatoria, y $\beta_1 \cdots \beta_k$ son los parámetros desconocidos a estimar por el modelo.

La estimación de los parámetros de este modelo implica la utilización de variables reales que midan la relación funcional definida. La búsqueda de las variables medibles no es asunto fácil ya que por una parte, la teoría no especifica cuál variable de la contabilidad nacional debe ser utilizada y, por otra parte, la estadística económica disponible no es generada bajo un plan y objetivos de análisis económico, es decir no es controlada por el economista y por ende no necesariamente se ajusta a sus necesidades de estudio de la realidad.

Los modelos econométricos pueden ser uniecuacionales o multiecuacionales. Los modelos uniecuacionales implican la estimación de una sola ecuación los multiecuacionales están formados por más de dos ecuaciones que pueden estar relacionadas entre sí. Los grandes modelos multiecuacionales han perdido importancia debido a la complejidad de su construcción y manejo, además de que el dominio metodológico de modelos más compactos, derivados de las propuestas VAR de formas reducidas, ha llevado a la utilización de modelos de pequeña escala. Sin embargo, aún se siguen actualizando modelos de gran escala para una amplia variedad de países debido a la necesidad de simulaciones de política que requieren los gobiernos, grandes empresas o bancos. Para el caso mexicano la empresa IHS sigue actualizando el primer modelo construido para el país en los años sesenta por CIEMEX una empresa asociada con la firma de modelos WARTHON Econometric Associates International. Actualmente ese modelo genera pronósticos de 800 variables para 25 sectores de la economía (IHS, 2013).

En el apartado anterior se argumentó que la metodología econométrica de libro de texto incorpora el supuesto de *"correcta especificación"* del modelo. La metodología moderna, al contrario, considera que las variables del modelo son aleatorias y por tanto sus propiedades probabilísticas son compartidas con el término de error.

Para formalizar esta idea consideremos el modelo de regresión como la media condicional de y_i sobre los valores de X_i:

$$FRP = E[y_i|X_{ji}] = f(X_{ji}) = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_{ki} \tag{2}$$

Donde:

- $j = 2, 3, \cdots, k$
- $i = 1, 2, \cdots, n$

A esta función se le conoce como función de regresión poblacional (FRP). La estimación de los parámetros de la función requiere de una regla que transforme las variables aleatorias en un estimador de los parámetros desconocidos.

La sustitución de los valores de una muestra particular de realizaciones de las variables aleatorias, en el estimador, genera una estimación de los parámetros desconocidos, la cual depende de la muestra y da lugar a una función de regresión muestral (FRM):

$$FRM = E[y_i|X_{ji}] = f(X_{ji}) = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \beta_k \hat{X}_{ki} \tag{3}$$

El término de error o innovaciones, a diferencia de la metodología tradicional, no es "añadido" a la función de regresión, se obtiene como la diferencia entre y_i y sumedia condicional:

$$[u_i|X_{ji}] = y_i - E[y_i|X_{ji}] = FIC \tag{4}$$

Que es conocida como la función de innovación condicional (FIC).

Así la ecuación para y_i puede escribirse como:

$$y_i = FRP + FIC \tag{5}$$

De esta manera la ecuación tendrá una parte sistemática que se corresponde con FRP y una no sistemática representada por FIC.

4. ECONOMETRÍA APLICADA Y R

El enfoque seguido en este texto es fundamentalmente de econometría aplicada, por ello se centra en las aplicaciones empíricas y se le brinda menor espacio a las discusiones teóricas y conceptuales. Es por lo tanto necesario contar con el manejo de paquetería computacional que permita la utilización de la metodología econométrica en una amplia variedad de métodos, datos reales y casos prácticos.

R es un lenguaje y un ambiente para manejo de datos y gráficos en código libre. Dada esas características los desarrollos que se han realizado en R son abiertos y están disponibles gratuitamente, por lo cual su uso se ha difundido ampliamente. R es difundido libremente por una gran diversidad de sitios espejo del **Comprehensive R Archive Network (CRAN)**. Además de ser gratuitas, los desarrollos para econometría en R se actualizan más rápido que en cualquier otro de los costosos softwares comerciales que se encuentran en el mercado. Esto es así debido a que los usuarios hacen desarrollos, los documentan y los suben al CRAN de R de manera cotidiana[4].

R genera objetos que son números, vectores, matrices, alfa numéricos y cuadros de datos. Los operadores aritméticos a los que usualmente estamos acostumbrados en otros paquetes son los mismos en R; suma (+), resta (-), multiplicación (*), división (/) y potencia (^). Los ejemplos siguientes están basados en Crawley (2009) y Venables et.al. (2013).

Por ejemplo, podemos generar un objeto número y que contiene el resultado de multiplicar 2 por 5:

```
# Creamos las dos variables con valores
a = 2
b = 5

# Creamos una variable donde multiplique ambas variables creadas
y = a*b
```

```
# Imprimimos el dato
y
```

```
Output Console
[1] 10
```

Los objetos que hemos creado los podemos listar con las siguientes opciones:

```
objects()
ls()
```

La ayuda se puede utilizar para obtener referencias de cualquier comando, por ejemplo si queremos sabes lo que hace la función `objects()`, basta escribir:

[4] R se puede descargar del siguiente vínculo:http://CRAN.R-project.org/

```
help("objects")
```

En seguida R despliega una ventana con toda la documentación del comando, en la cual noms brinda su descripción, uso, argumentos, detalles, referencias y ejemplos de su uso.

Los objetos pueden eliminarse rápidamente, por ejemplo para eliminar `a` y `b`, basta escribir el siguiente código:

```
# Eliminar variables creadas
rm(a, b)
```

Para generar un objeto que sea un vector columna, podemos usar la función `c()`:

```
x = c(5, 10, 8, 7, 9)
```

Lo mismo puede hacerse con la función `assign()`:

```
assign("x", c(5, 10, 8, 7, 9))
```

Es posible calcular la media, `mean()`, la varianza, `var()`, el valor máximo, `max()`, el valor mínimo, `min()`, o la longitud del vector, `length()`. Por ejemplo, si calculamos la media.

```
mean(x)
```

```
Output Console
[1] 7.8
```

También podríamos generar vectores columna con secuancias de números, por ejemplo si generamos una secuancia del 1 al 10:

```
y = c(1:10)

y
```

```
Output Console

[1]  1  2  3  4  5  6  7  8  9 10
```

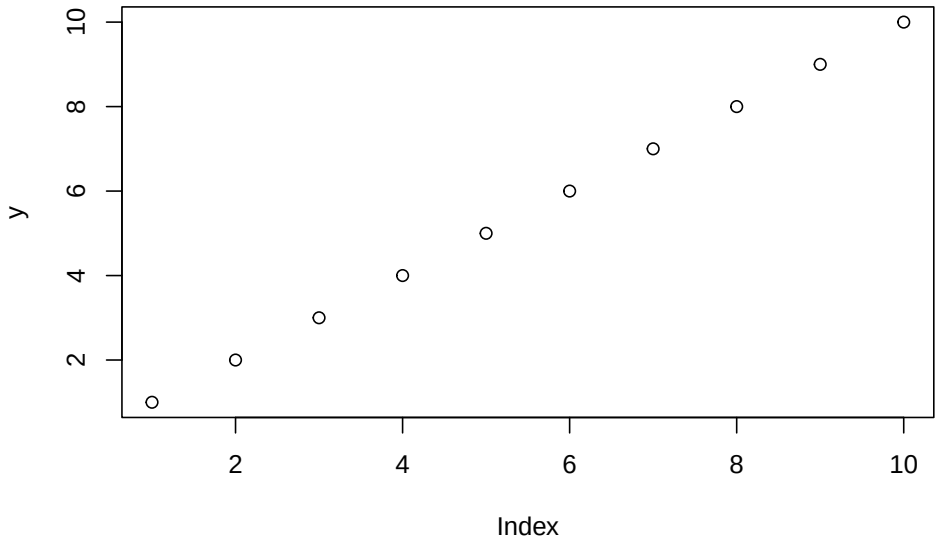
A los elementos de un vector se les puede asignar nombres, por ejemplo, al vector `x` le asignaremos los nombres de los números que contiene:

```
names(x) = c("cinco", "diez", "ocho", "siete", "nueve")

# Imprimimos la variable
x
```

```
Outpur Console
cinco  diez  ocho siete nueve
    5    10     8     7     9
```

Las gráficas se obtienen usando la función `plot()`, por ejemplo, para realizar una gráfica de los valores del vector y escribimos:

```
plot(y)
```



Con el fin de ejemplificar algunas opciones que se utilizarán ampliamente al estimar modelos de regresión vamos a considerar el caso siguiente. Generamos dos vectores con la siguiente información:

```
y = c(1,2,3,-1,0,-1,2,1,2)
x = c(0,1,2,-2,1,-2,0,-1,1)
```

Ahora es posible correr la regresión para el modelo $y_i = \beta_1 + \beta_2 X_{2i} + u_i$. Por el momento no se preocupe de las características del modelo, ni de la comprensión del método de estimación, ya que eso aborda en los capítulos siguientes del libro. Aquí simplemente debe aprender que a correr esa regresión se utiliza la función lineal model o `lm()`:

```
lm(y ~ x)
```

```
Output Console
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     1.0000       0.8125
```

Los resultados de la regresión se pueden obtener con la función `summary()`:

```
summary(lm(y ~ x))
```

```
Output Console
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8125 -0.3750  0.1875  0.3750  1.0000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0000     0.2938   3.404  0.01138 *
x             0.8125     0.2203   3.688  0.00778 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8814 on 7 degrees of freedom
Multiple R-squared:  0.6602,    Adjusted R-squared:  0.6116
F-statistic:  13.6 on 1 and 7 DF,  p-value: 0.007782
```

Ahora ya estammos en condiciones de preparar nuestros datos para utilizarlos en el paquete. La manera más fácil de manejar los archivos de datos en R es crearlos en una hoja de cálculo en Excel y así guardarlos como archivo de texto delimitado por tabuladores.

Los datos del archivo **Cap1_Ejercicio1.csv** fueron guardados en formato delimitado por comas. En el archivo se presentan los datos de la muestra de un país hipotético, como el **PIB**, la inversión (**FBKF**) y sus **Años** consecutivos, desde 1900 hasta 2021.

Para abrir el archivo en R, primero se tiene que asegurar que el paquete esté direccionando a la carpeta en el que ha guardado el archivo. Para verificar cuál es el directorio actual de trabajo, sólo se escribe la función `getwd()`.

Si el sirectorio que aparece no es el que se debe de utilizar, se puede cambiar de directorio con la función `setwd("C:\Users\Directorio_a_usar")`.

Para que los datos puedan ser cargados en `R`, se debe usar la función `read.xlsx()`. Para ello, tendremos que instalar el paquete `openxlsx`, de la forma que se muestra a continuación.

```
#  Paqueterías a instalar
install.packages("openxlsx")
library(openxlsx)

# Ruta del documento en donde se encuentra el archivo
datos = read.xlsx("C:\\Users\\CAPÍTULO 1\\Cap1_Ejercicio1.xlsx")
```

Ahora, al pedir un listado en `R` aparecerá cada una de las variables de la lista:

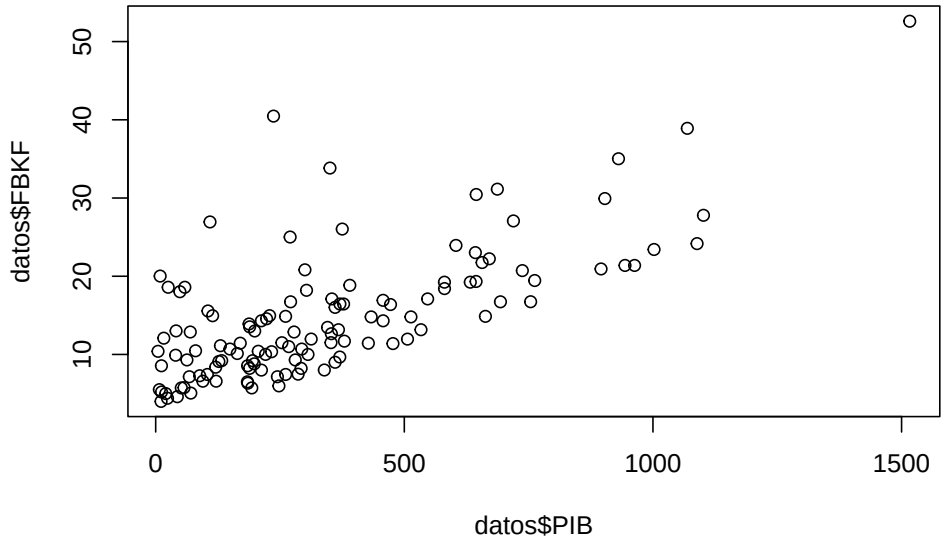
```
# Pedimos el listado de variables
ls(datos)
```

```
## [1] "Año"  "FBKF" "PIB"
```

```
Output Console
[1] "Año"  "FBKF" "PIB"
```

Una herramienta gráfica que utilizaremos frecuentemente es un **diagrama de dispersión**. Por ejemplo, se puede solicitar un diagrama de dispersión para visualizar la relación entre el **PIB** y la **FBKF**.

```
plot(datos$PIB, datos$FBKF)
```



En el diagrama se puede observar claramente una relación positiva entre el Producto Interno Bruto y La Inversión de un país hipotético de la muestra de datos.

Como ya sabemos utilizar el comando de regresión, podemos ahora estimar un modelo para explicar el Producto Interno Bruto del país en función a la Inversión, pero ahora guardaremos el resultado en un objeto nombrado 'regresion_1".

```
# Generamos el modelo de regresión lineal
regresion_1 = lm(PIB ~ FBKF, data = datos)
```

Los resultados del modelo indican que al incrementar en una unidad la Formación Bruta de Capital Fijo, el PIB incrementa en un 25.16 unidades, tal y como se aprecia en los siguientes resultados.

Para observarlo se usa la función `summary()` y dentro nuestra variable nombrada como 'regresion_1".

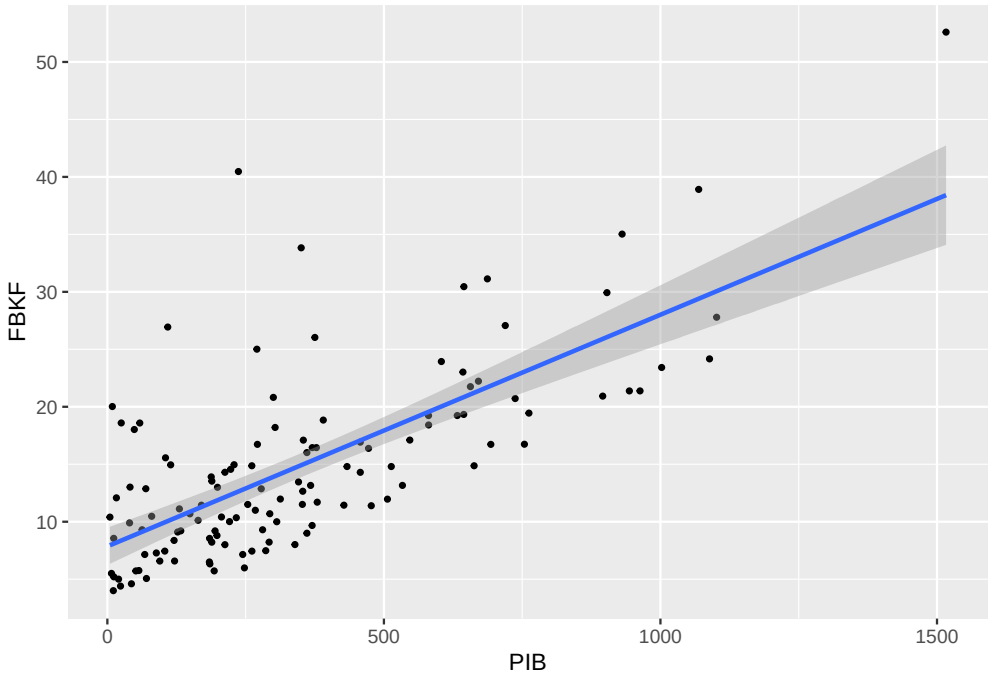
```
# Observamos el resumen del modelo
summary(regresion_1)
```

```
# Resumen del modelo
Output console
Call:
lm(formula = PIB ~ FBKF, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-754.67  -94.45   -6.64  121.55  507.19

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -26.568     38.439  -0.691    0.491
FBKF          25.160      2.265  11.110   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 205.7 on 120 degrees of freedom
Multiple R-squared:  0.507, Adjusted R-squared:  0.5029
F-statistic: 123.4 on 1 and 120 DF,  p-value: < 2.2e-16
```

Para añadir la recta dde regresión al diagrama de dispersión, se puede usar el paquete `ggplot2` de la siguiente manera:

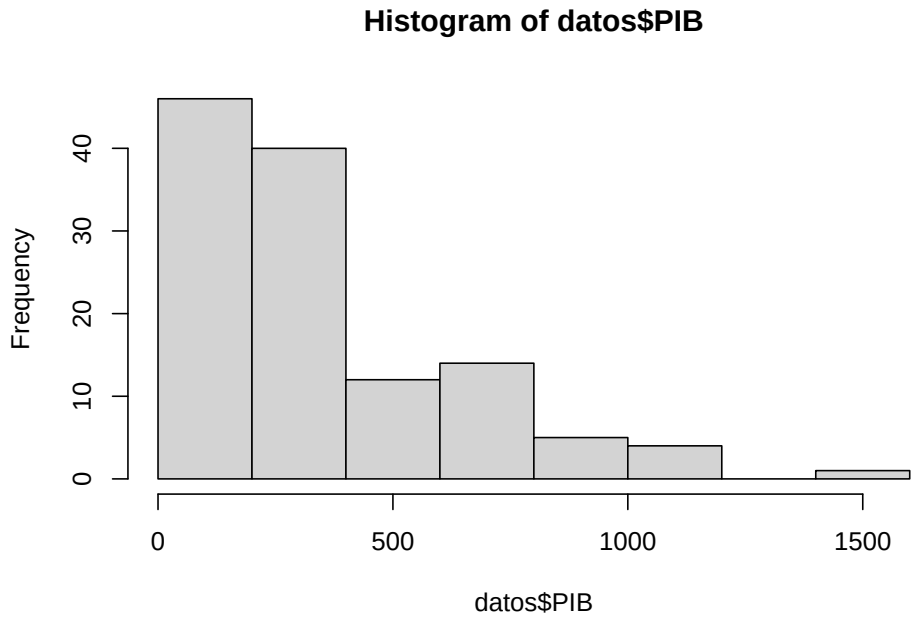
```
# Instalamos el paquete
install.packages("ggplot2")
library(ggplot2)
```

```
# Generamos el diagrama de dispersión con la recta de regresión
ggplot(data = datos, aes(
  x = PIB, y = FBKF
)) +
  # Añadimos los puntos de dispersión
  geom_jitter(size = 0.9) +
  # Añadimos la recta de regresión
  geom_smooth(method = "lm")
```



Otro diagrama que llega a ser de utilidad es el **histograma**. El histograma permite relacionar intervalos de los datos de frecuencia. Con la función `hist()` se puede generar el histograma para los datos del PIB:

```
hist(datos$PIB)
```

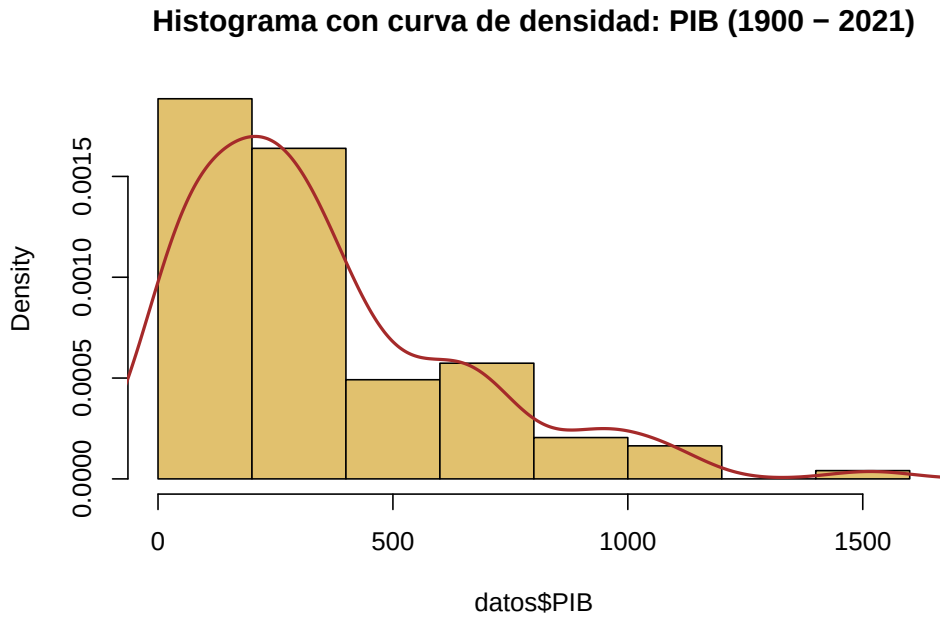


Claramente el histograma muestra que la mayor frecuencia de datos del PIB se encuentra con loa ingresos menores de 500, siendo los más bajos de la distribución.

Resulta algo útil visualizar el histograma en densidades (el área bajo la curva igual a la unidad), y así añadirle funciones de densidad Kernel, lo cual se puede visualizar con la función `hist()` y añadir la función `lines()` para agregar la curva de densidad[5].

```
# Se crea el Histograma
hist(datos$PIB, freq = FALSE,
     col = "#E1C16E",
     main = "Histograma con curva de densidad: PIB (1900 - 2021)")

# Se agrega la curva de densidad
densidad = density(datos$PIB)
lines(densidad, col = "#A52A2A", lwd = 2)
```

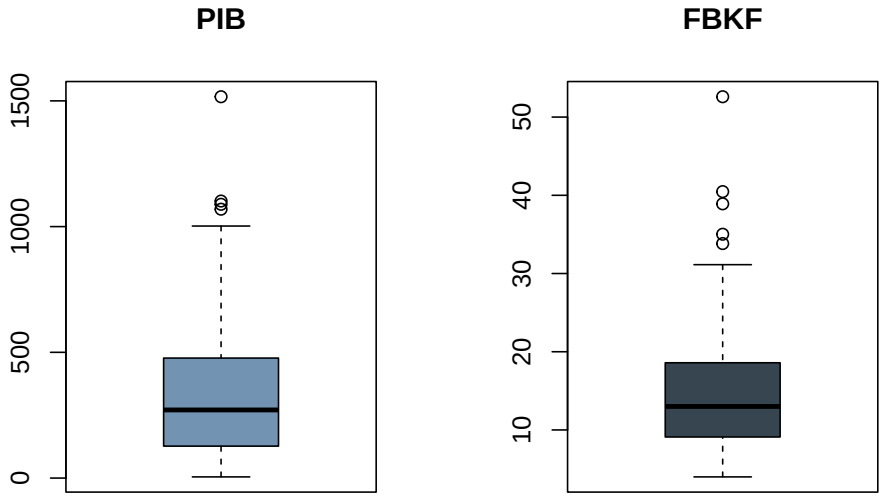
[5]Usando la función `hist()` se crea el histograma. La opción de `freq = FALSE` se utiliza para que el histograma muestre la densidad de las frecuencias. La opción `col = #E1C16E` se utiliza para personalizar el color de las barras del histograma. La función `main = "Histograma con curva de densidad: PIB (1900 - 2021)"` se utiliza para agregar un título al diagrama. Se utiliza también la función `lines()` para agregar la curva de densidad al diagrama.



Otra forma de observar la distribución de los datos es utilizar un diagrama de cajas (**boxplot**), en las cuales el diagrama muestra los umbrales para los cuartiles inferior y superior, además de la mediana. Las líneas inferiores y superiores de la caja permiten identificar las observaciones extremas. Para obtener este tipo de diagrama se usa la función `boxplot()`, como se muestra a continuación[6]:

```
par(mfrow = c(1,2))
# Diagramas de cajas

# Para el PIB
boxplot(datos$PIB, col = "#7393B3",
        main = "PIB")
# Para el FBKF
boxplot(datos$FBKF, col = "#36454F",
        main = "FBKF")
```

[6]Para observar dos diagramas al mismo tiempo, se usa la función `par(mfrow = c(1,2))`, y se regresa a la función `par(mfrow = c(1,1))` para que los siguientes gráficos sean puestos en el mismo lugar.



5. ALGUNOS DESARROLLOS EN R QUE FACILITAN EL USO DE LA ECONOMETRÍA

En R se cuenta con interfaces que pueden llegar a utilizar de forma más amigable los recursos disponibles en este software. Una de estas interfaces (IDE: Integrated Development Environment para las siglas en inglés de Entorno de Desarrollo Integrado) es RStudio, la cual se puede instalar desde el siguiente vínculo:

- https://www.rstudio.com

La primer ventaja de RStudio es que permite visualizar los datos y su historial de trabajo en la ventanda de **WORKSPACE / HISTORY**, al mismo tiempo es posible ver la ventana **CONSOLA** en la cual se ejecutan los comandos de R, cuenta también con una ventana en la cual se pueden visualizar los datos de ayuda, **HELP**, archivos, **FILE**, diagramas, **PLOT**, y paqueterías, **PACKAGES**. La cuarta ventana es el **SPURCE** en la que se muestran los archivos de origen.

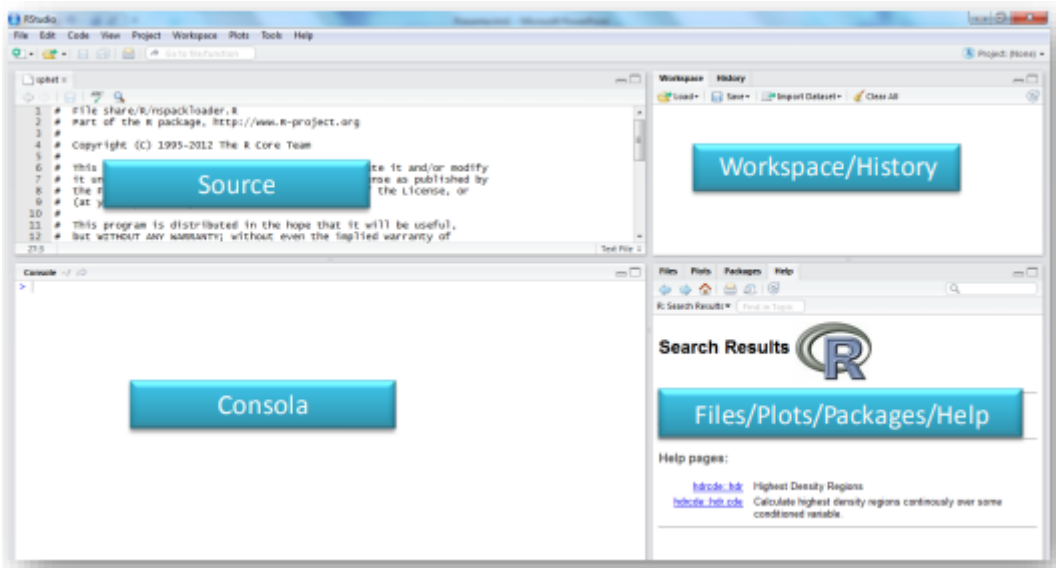


Figura 2: Interface de RStudio

REFERENCIAS

- Crawley, J. Michael (2009), The R book, ed. Wiley, Inglaterra.
- Fox, John (2005), The R Commander: A Basic-Statistics Graphical User Interface to R, Journal of Statistical Software, vol.14, núm. 9, pp. 1-42.
- Hoover D., Kevin (2006), The methodology of econometrics, en Terence Mills y Kerry Patterson, Plagrave Handbook of Econometrics, vol.1, Econometric Theory, Palgrave Mcmillan, pp. 61-87, Reino Unido.
- Maddala, G. S. (1996). Introducción a la econometría. Ed. Prentice Hall, México.
- Spanos, Aris (1996). Statistical Foundation of econometric modeling. Ed. Cambridge University Press.
- Spanos, Aris (2006), Econometrics in retrospect and prospect, en Terence Mills y Kerry Patterson, Plagrave Handbook of Econometrics, vol.1, Econometric Theory, Palgrave Mcmillan, pp. 3-58, Reino Unido.
- Venables, W. N. y D. M. Smith (2013), An introduction to R, ed. R Core Team.

REFERENCIAS ELECTRÓNICAS

- CRAN (2021), https://www.r-project.org/
- IHS (2021), https://www.spglobal.com/marketintelligence/en/mi/industry/economics-country-risk.html

- Penn Tables (2021), https://pwt.sas.upenn.edu/
- RStudio (2021), https://posit.co/download/rstudio-desktop/

ARCHIVOS DE DATOS ASOCIADOS AL CAPÍTULO

- CAPÍTULO 1

CAPÍTULO 2: ENFOQUE MATRICIAL DE LA REGRESIÓN LINEAL

- Por Galán Figueroa, Javier.

1. EL MODELO MATRICIAL

En este capítulo se considera relevante que el usuario conozca, en primera instancia, las rutinas básicas que son necesarias para estimar los parámetros de la regresión lineal a través del enfoque matricial, utilizando la paquetería del software `R`, los cuales podrán ser utilizados en sus variantes como es el `RStudio`.

Para comenzar, se utilizarán datos de la economía mexicana para el periodo de Enero de 1993 a Diciembre del 2002, con frecuancia mensual, cuya fuente proviene de la página web de INEGI. Con dicha información permitirá estimar el siguiente modelo:

$$y = f(X_2, X_3) \tag{6}$$

$$y = X\beta + u \tag{7}$$

$$y_t = \beta_1 + \beta_2 X_{2t} + \cdots + \beta_i X_{3i} + u \tag{8}$$

La ecuación (7) es la representación matricial de la regresión lineal, donde y es un vector columna de orden $(nx1)$, X es una matriz de orden (nxk), β es un vector columna de orden $(kx1)$, por último u es un vector columnna de orden $(nx1)$, es decir:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \tag{9}$$

De la ecuación (8) la variable dependiente, y, es el nivel de Exportaciones de Bienes y Servicios Finales[7] que es explicada por el nivel de Producto Interno Bruto[8], X_2, y por el Consumo Privado[9], X_3.

Para encontrar el modelo en el cual explique el comportamiento del Niveer de Exportaciones de Bienes y Servicios Finales en función del Producto Interno Bruto y por el Consumo Privado, se utilizará los datos que se encuentran dentro del archivo **EJECRICIO_2.xlsx**. Para ejecutarlo en `R` se hce uso de la siguiente forma usando la librería `openxlsx` con la función `read.xlsx`:

```
# Librería a cargar
#install.packages("openxlsx")
library(openxlsx)
```

[7] En millones de Pesos Mexicanos. Deflactados a precios del año base 2013 = 100.
[8] En millones de Pesos Mexicanos. Deflactados a precios del año base 2013 = 100.
[9] En millones de Pesos Mexicanos. Deflactados a precios del año base 2013 = 100.

```
# Se carga la ruta del archivo y se nombra como se desea
Ejercicio_2 = read.xlsx(("C:\\Users\\CAPÍTULO 2\\EJECRICIO_2.xlsx")
```

Si se desean visualizar los datos a través de una lista, basta con escribir nuestra variable nombrada dentro de la función `head()`. Esto permitirá visualizar los primeros 5 datos de nuestras variables por columnas que se tienen.

```
# Ver los datos
head(Ejercicio_2)
```

```
Output Console
Periodos       X       G       C    TOTAL      PIB      IM
1    33970 1421049 1410454 6094268 11467308 10008895 1458413
2    34001 1477721 1428917 6304694 11745542 10171035 1574506
3    34029 1465346 1376521 6243679 11624474 10066258 1558216
4    34060 1588556 1423995 6598255 12133939 10416096 1717843
5    34090 1576503 1453657 6258159 12072447 10343388 1729058
6    34121 1594674 1483317 6643072 12619228 10772526 1846702
```

2. ANÁLISIS EXPLORATORIO DE LOS DATOS

Después de haber cargado los datos al programa, se procederá a realizar el siguiente análisis estadístico de las variables.

Si se desea obtener de manera individual los siguientes parámetros: Media aritmética, Mediana, Desviasión estándar y varianza, sólamente se usan las siguientes funciones, en el orden anteriormente nombrado:

- `mean()`
- `median()`
- `sd()`
- `var()`

De manera conjunta se usa la función `summary()`.

Veamos la función en su utilidad por parte de la variable de las Exportaciones de Bienes y Servicios Finales, X.

```
# Análisis exploratorio Individual
mean(Ejercicio_2$X)
median(Ejercicio_2$X)
sd(Ejercicio_2$X)
var(Ejercicio_2$X)
```

```
Output Console
[1] 4253964
[1] 3991206
[1] 1714253
[1] 2.938663e+12
```

```
# Resumen general
summary(Ejercicio_2$X)
```

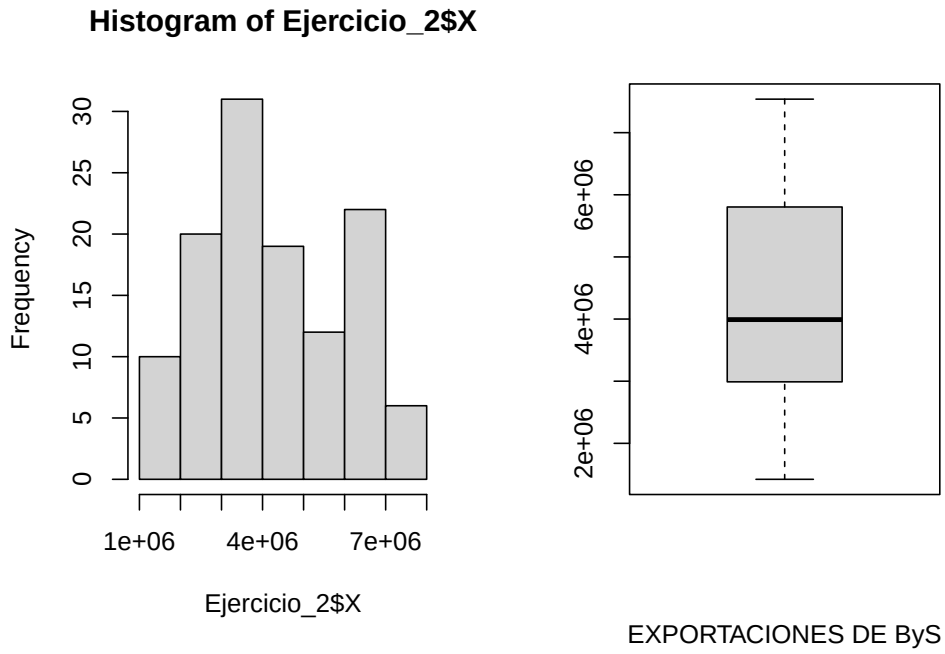
```
Output Console
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1421049 3002629 3991206 4253964 5760310 7539036
```

Entonces, podemos analizar que, para el nivel de Exportaciones de Bienes y Servicios Finales:

- El mínimo es de $1,421,049$ MXN.
- El máximo es de $7,539,036$ MXN.
- El promedio es de $4,253,964$ MXN.

De lo anterior, R agrupa los datos y calcula los cuartiles. El primero es de $3,002,629$ MXN, mientras que el segundo cuartil (o mediana) es de $3,991,206$, y el tercer cuartil es de $5,760,310$. Posteriormente se puede visualizar su histograma y su diagrama de caja.

```
# Dividir en 2 el panel
par(mfrow = c(1, 2))

hist(Ejercicio_2$X)
boxplot(Ejercicio_2$X, sub = "EXPORTACIONES DE ByS")
```

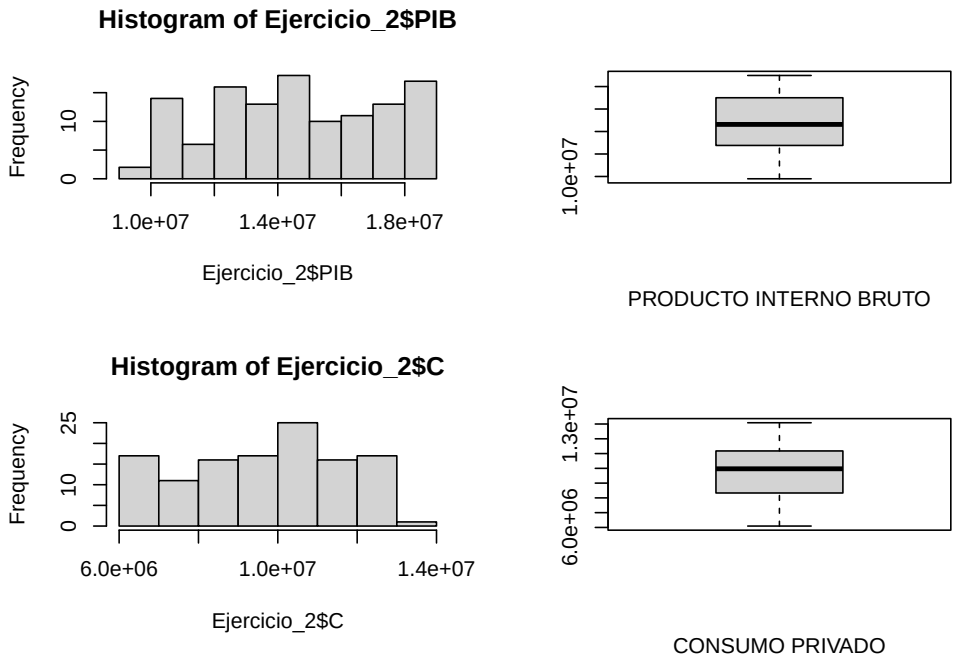


Se puede repetir el mismo orden anterior para las variables independientes del modelo, PIB, C:

```
# Resumen de las variables dependientes
summary(Ejercicio_2$PIB, Ejercicio_2$C)
```

```
Output Console
 Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 9795719 12778345 14631966 14596128 16922340 18985339
```

```r
par(mfrow = c(2, 2))

# Para el PIB
hist(Ejercicio_2$PIB)
boxplot(Ejercicio_2$PIB, sub = "PRODUCTO INTERNO BRUTO")

# Para el C
hist(Ejercicio_2$C)
boxplot(Ejercicio_2$C, sub = "CONSUMO PRIVADO")
```



A continuación se utilizará la función `cor()` para obtener la **matriz de correlación** entre las variables (y, X_2, X_3):

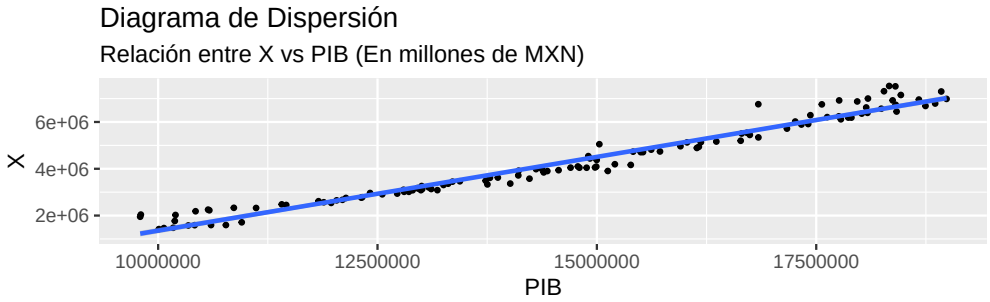
```r
# Generamos un nuevo dataframe con las variables a usar
Datos_2 = data.frame(Ejercicio_2$X,
                     Ejercicio_2$PIB,
                     Ejercicio_2$C)

# Se renombran los nombres de las columnas
colnames(Datos_2) = c("X", "PIB", "C")

# Matriz de correlación
cor(Datos_2)
```

```
Output Console
            X       PIB         C
X   1.0000000 0.9820784 0.9647438
PIB 0.9820784 1.0000000 0.9894907
C   0.9647438 0.9894907 1.0000000
```

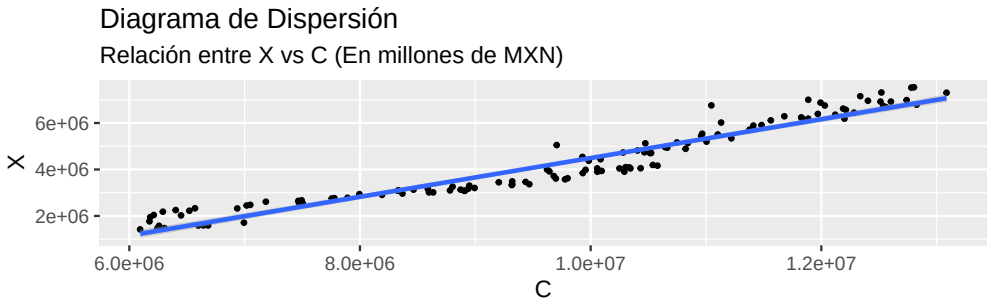
De acueerdo a la matriz de correlación, la asociación entre las variables (X_2, y) es positiva y del 0.9820784 o del 98.20 %. Mientras que la asociación entre (X_3, y) es de igual manera positiva y del 0.9647438 o del 96.47 %. Por otro lado, las variables (X_2, X_3) se asocian en 0.9894907 o en 98.94 %.

Para obtener los diagramas de dispersión para indicar a nivel gráfico cómo influye el Producto Interno Bruto y el Consumo Privado al nivel de Exportaciones de Bienes y Servicios Finales, se prosigue con usar la librería `ggplot2` para generar un diagrama de dispersión a través de la función `geom_jitter()` y la función `geom_abline()` para la línea regresora, argumentando `method = "lm"` para que sea función la línea de regresión.

```
install.packages("ggplot2")
library(ggplot2)

# X vs PIB
ggplot(Datos_2, aes(x = PIB, y = X)) +
  geom_jitter() +
  geom_smooth(method = "lm")

# X vs C
ggplot(Datos_2, aes(x = C, y = X)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```



Fuente: Elaboración propia a partir de datos INEGI (2023)



Fuente: Elaboración propia a partir de datos INEGI (2023)

3. ESTIMACIÓN POR MÍNIMOS CUADRADOS ORDINARIOS (MCO)

Con el análisis previo se procederá a estimar los parámestros de la ecuación (8) a través de Mínimos Cuadrados Ordinarios (MCO)[10]. Para ello, se considera que el vector β de la ecuación (7) es estimable a partir de la siguiente expresión:

$$\beta = (X'X)^{-1}X'y \tag{10}$$

Como primer paso se debe especificar en el programa `R` la matriz X, así como el vector y. Para ello, se sigue el siguiente algoritmo:

1. Para transformar un conjunto de variables a la matriz se usa la función `cbind()`.
2. Una vez que se ha dado de alta las matrices en `R` se procede a realizar las operaciones correspondientes paara encontrar los componentes del vector $(X'X)^{-1}X'y$, los cuales se describen a continuación.

Para crear la matriz X, que conforma de acuerdo a la ecuación (9), se utiliza el siguiente código:

```
X = cbind(1, Datos_2$PIB, Datos_2$C)
```

Donde las opciones que aparecen dentro del paréntesis indican que el uno hace referencia al intercepto, mientras que las otras variables son las independientes. Para el caso a transformar la variable dependiente, se usará el siguiente código:

```
y1 = cbind(Datos_2$X)
```

Para estimar el vector β de la ecuación (5), primero se obtiene el producto $(X'X)$ para ello se siguen lo siguientes pasos:

[10]El método de mínimos cuadrados ordinarios (MCO) es una técnica utilizada en estadística y análisis de datos para encontrar la mejor línea de ajuste a través de un conjunto de puntos de datos en un modelo de regresión lineal. El objetivo del MCO es encontrar la línea de regresión que minimiza la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo. En otras palabras, el método de MCO encuentra la línea de regresión que mejor se ajusta a los datos disponibles, minimizando la suma de los cuadrados de los residuos, que son las diferencias entre los valores observados y los valores predichos por el modelo. El método de MCO se utiliza en muchos campos, incluyendo la econometría, la ingeniería, la física, la biología y la ciencia social. Es una técnica muy útil para modelar la relación entre dos variables, especialmente cuando se sospecha que existe una relación lineal entre ellas. En `R` y `Python`, el método de MCO se puede implementar utilizando diversas funciones, como `lm()` en `R` y `LinearRegression()` en `Python`, entre otras. Estas funciones permiten ajustar una línea de regresión lineal y calcular sus coeficientes y estadísticas asociadas, como el R-cuadrado, el error estándar y el intervalo de confianza.

1. Transpuesta de X
2. Producto de la trnaspuesta de X por X

Cabe mencionar que `R` puede llevar el producto de matrices mediante la función `%*%`.

```
# Transpuesta de X
trX = (t(X))

# Producto de la transpuesta de X por X
X_X = trX %*% X

# Resultado
X_X
```

```
Output Console
            [,1]          [,2]          [,3]
[1,]         120 1.751535e+09 1.165471e+09
[2,] 1751535376 2.641389e+16 1.763422e+16
[3,] 1165470709 1.763422e+16 1.178648e+16
```

A continuación, se obtiene el determinante de la matriz $(X'X)$, para determinar si ésta tiene inversa o no. Para obtener la inverta, $(X'X)^{-1}$, se debe primero activar la librería `MASS`, después usar la función `ginv()`.[11]

```
# Se obtiene el determinante de X_X
det(X_X)

# Instalar la librería MASS
# install.packages("MASS")
library(MASS)

# Se genera la inversa de X_X
invX_X = (ginv(X_X))

# Obtenemos los datos de la inversa de X_X
invX_X
```

[11]La librería `MASS` (Modern Applied Statistics with S) en `R` es una de las librerías más utilizadas en el campo de la estadística y el análisis de datos. Esta librería proporciona una amplia gama de funciones y herramientas para realizar análisis estadísticos y modelos de regresión avanzados. Algunas de las funciónes más comunes incluyen: **1)** Análisis de Componentes Principales (PCA) y Análisis Discriminante Lineal (LDA). **2)** Modelos Lineales Generalizados (GLM), incluyendo modelos de regresión logística y Poisson. **3)** Métodos no paramétricos como regresión spline y suavizado por kernel. **4)** Modelos de mezcla finita, incluyendo modelos de mezcla de Gaussianas y análisis de conglomerados. La librería `MASS` es particularmente útil para aquellos que trabajan en ciencias sociales, económicas y médicas, así como en campos de investigación relacionados con la bioestadística, la bioinformática y la genómica.

```
Output Console
              [,1]          [,2]          [,3]
[1,]  1.510673e-26  2.220737e-20 -3.324940e-20
[2,]  2.220734e-20  3.266024e-14 -4.886431e-14
[3,] -3.324936e-20 -4.886431e-14  7.319272e-14
```

Una vez que se tiene la matriz inversa, $X'X^{-1}$, se procede a obtener el producto de $X'y$:

```
# Producto de Xy
Xy = trX %*% y1

# Resultado de Xy
Xy
```

```
Output Console
             [,1]
[1,] 5.104757e+08
[2,] 7.985852e+15
[3,] 5.347792e+15
```

Por último, se procede a calcular el vector beta, β, a través del la siguiente forma de código:

```
# Se calcula beta
beta = invX_X %*% Xy

# Se ven los resultados de beta
beta
```

```
Output Console
              [,1]
[1,] -4.661530e-07
[2,] -4.963258e-01
[3,]  1.196296e+00
```

Un método de comprobación para obtener la certeza que este vector, el cual fue obtenido paso a paso mediante **álgebra lineal**, se utiliza el código para estimar de manera directa el modelo de regresión lineal, `lm(y ~ x)`, cabe mencionar que `R` utiliza el mismo método.

```
# Se estima el modelo de regresión lineal
modelo = lm(X ~ PIB + C, data = Datos_2)
```

```
# Resumen del modelo
summary(modelo)
```

```
Output Console
Call:
lm(formula = X ~ PIB + C, data = Datos_2)

Residuals:
    Min      1Q  Median      3Q     Max
-626093 -176378  -75129   83301 1031266

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.242e+06  1.897e+05 -27.631  < 2e-16 ***
PIB          8.437e-01  7.477e-02  11.284  < 2e-16 ***
C           -2.902e-01  1.008e-01  -2.881  0.00472 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 314900 on 117 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9663
F-statistic:  1705 on 2 and 117 DF,  p-value: < 2.2e-16
```

REFERENCIAS

- Crawley, Michael (2013), *The R Book*, 2a. Ed., Wiley, United Kingdom.
- Green, William (2003), *Econometric Analysis*, 5a Ed., Pearson Education. EUA.
- Johnston, J. y J. Dinardo (1997), *Econometrics Methods*, 4a Ed., McGraw-Hill. EUA.
- Quintana, L. y M. A. Mendoza (2008), *Econometría Básica. Modelos y aplicaciones a la economía mexicana*, Plaza y Valdés Editores, México.

ARCHIVOS DE DATOS ASOCIADOS AL CAPÍTULO

- CAPÍTULO 2

CAPÍTULO 3: EL MODELO DE REGRESIÓN MÚLTIPLE

- Por Feregrino Feregrino, Jorge.

1. ESPECIFICACIÓN DEL MODELO DE REGRESIÓN MÚLTIPLE

El primer paso en la especificación de un modelo econométrico es identificar el objeto de investigación en relación al área de estudios de las ciencias socioeconómicas. En esta etapa, es necesario recopilar información acerca del comportamiento teórico del objeto de investigación para identificar patrones de comportamiento, situar alguna problemática específica y plantear las hipótesis necesarias. La especificación del modelo nos permitirá explorar las hipótesis principales, identificar las relaciones que explican el objeto de estudios y diseñar una propuesta teórica alternativa de acuerdo a los objetivos del usuario.

La identificación del objeto de investigación permitirá realizar una búsqueda exhaustiva de los datos para llevar a cabo una aproximación del comportamiento del fenómeno mediante los hechos estilizados. Una vez identificada la problemática se procede a establecer las relaciones y la selección de las variables. La búsqueda de la información de las variables, la relación teórica y la descripción estadística de estas será útil para determinar la metodología de análisis. En el caso de la mayoría de los hechos socioeconómicos los fenómenos están determinados por un conjunto de variables que puede llegar a ser infinito.

En economía se pueden identificar diversas relaciones teóricas entre variables; por ejemplo la producción para la teoría neoclásica está determinada por la combinación entre capital y trabajo, en la teoría keynesiana el ingreso de una economía cerrada está determinado por el consumo, la inversión y el gasto de gobierno, la tasa de inflación se puede determinar por la brecha del producto y las expectativas de inflación dentro del esquema de metas de inflación; así los ejemplos anteriores representan algunas de las problemáticas que se resuelven a través del establecimiento de relaciones entre variables.

En los modelos econométricos se establecen a priori las relaciones funcionales, con los elementos que se han descrito, para identificar los vínculos fundamentales entre las variables seleccionadas. De esta forma, se establecen las variables independientes y las dependientes. La elección de la variable dependiente y las independientes conformarán una relación funcional múltiple para describir el fenómeno económico mediante la metodología econométrica propuesta.

En el modelo de regresión múltiple las variables exógenas (X_j), asociadas a coeficientes lineales constantes (β_j), indican el efecto condicionado de cada variable independiente sobre la variable dependiente (Y), la especificación general del modelo con cuatro variables independientes es la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \tag{11}$$

Por ejemplo: El administrador de una tienda quiere determinar los mejores criterios para elegir la localización de algunas tiendas, una de las primeras sugerencias para la especificación del modelo es elegir la variable dependiente en este caso serían las ventas:

$$Y = Ventas \tag{12}$$

Posteriormente, se realiza la recomendación sobre la elección de las variables independientes, en este caso la teoría plantea que múltiples variables inciden en el comportamiento de las ventas (Y), se consideran las siguientes:

- X_1: Tamaño de la tienda
- x_2: Tráfico de personas en la calle
- x_3: Tiendas rivales en la zona
- x_4: Renta per cápita de la población residente en la zona
- x_5: Número total de personas que residen en la zona

La especificación sería una forma funcional lineal, donde se busca encontrar el grado de relación entre la variable **endógena**, Y, con las variables **exógenas**, $X_1, X_2, X_3, \cdots, X_5$. La forma funcional en la mayoría de los modelos, debe incorporar los **errores** que se generan en la estimación de la relación funcional entre las variables. La relación entre las variables es inexacta, por lo tanto, la evaluación se realiza en términos probabilísticos.

Ejercicio en `R`: Retomando el ejemplo de localización de tiendas y a fin de estimar el modelo de regresión se debe importar la base de datos a la cual se asignará el nombre *"tiendas"*, a la columna de datos de la variable dependiente se le asignará el nombre *"ventas"*, mientras que los nombres de las variables independientes quedarán de la siguiente forma:

- X_1: "tamaño"
- X_2: "tráfico"
- X_3: "rivales"
- X_4: "renta"
- X_5: "residentes"

La función para importar los datos desde Excel es `read.delim()`.

```
read.delim("ruta de acceso", sep = ",",
           header = TRUE, stringsAsFactors = FALSE)
```

La forma funcional reducida de la estimación de la regresión múltiple, al expresarse en términos probabilísticos debe incorporar un término de error (ε_i):

$$\hat{y}_i = b_0 + \Sigma_{j=1}^{k} b_j x_{ji} + \varepsilon_i \tag{13}$$

La estimación de una regresión múltiple tiene los siguientes objetivos:

1. Estimar los valores de una variable independiente ($\hat{y}$) mediante una función lineal de un número (K) variables independientes observadas x_j, donde: $j = 1, \cdots, K$. La representación es la siguiente:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{kn} \tag{14}$$

Donde: $i = 1, \cdots, n$ observaciones.

2. Obtener los efectos estadísticos de cada variable independiente, mediante la estimación de los coeficientes b_j, sobre la variable dependiente $(\hat{y})$. El coeficiente b_j de cada variable dependiente indica el impacto que tiene una variación unitaria de x_j, descontando el efecto simultaneo que tienen las otras variables independientes, es decir, se mantiene la independencia entre estas variables.

3. Estimar la exogeneidad débil, para mostrar que la distribución marginal de la variable independiente, al no contener información relevante para estimar los parámetros de interés, se puede eliminar.

El modelo de regresión múltiple poblacional sería el siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \tag{15}$$

El modelo de regresión múltiple permite obtener estimaciones simultáneas de b_j a partir del modelo poblacional β_j.

2. ESTIMACIÓN DE LOS COEFICIENTES DE REGRESIÓN

La estimación de la forma funcional múltiple, parte de los siguientes supuestos sobre los coeficientes a obtener:

1. Las variables independientes, x_{ji}, son números fijos, o bien, variables aleatorias, X_j, independientes del término de error, ε_i.
2. El valor esperado de la variable aleatoria, $\hat{y}$, es una función de las variables independientes, X_j.
3. Los términos de error estocásticos, ε_i, son variables cuya medida esperada es igual a cero, y la varianza constante, σ^2, para todas las observaciones:

- $E[\varepsilon_i] = y$
- $E[\varepsilon_i^2] = \sigma^2$

Para todo $i = 1, 2, \cdots, n$

4. Los términos de error aleatorios, ε_i, no están correlacionados entre sí:

- $E[\varepsilon_i \varepsilon_j] = 0$

Para todo $i = j$

5. No es posible hallar un conjunto de números que no sean iguales a cero, tal que:

- $c_0 + c_1 x_{1i} + c_2 x_{2i} + \cdots + c_k x_{ki} = 0$

Esto probaría la ausencia de relación lineal entre las variables X_j. Los primeros 4 supuestos están implícitos en la regresión simple, el quinto supuesto excluye cualquier posibilidad de relación lineal entre las variables independientes, y nos permite hacer una selección específica de las variables y su impacto sobre la variable independiente en una regresión múltiple.

El método utilizado para estimar los coeficientes de la regresión múltiple es el de **Mínimos Cuadrados Ordinarios** (MCO), los coeficientes se obtienen mediante la minimización de los errores o la suma de los residuos explicados al cuadrado, SCE. En un primer modelo, los errores en el tiempo están explicados por las desviaciones de las variables independientes observadas, y_i, en el tiempo con relación a la variable explicada, $\hat{y}_i$:

- $e_i = y_i - \hat{y}_i$

Para minimizar la SCE se procede de la siguiente forma matemática. La SCE tiene la siguiente representación:

$$SCE = \Sigma_{i=1}^{n} e_i^2 = \Sigma_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{16}$$

De la sumatoria se extraen las diferencias elevadas al cuadrado entre los valores de y_i y los valores de la variable estimada, $\hat{y_i}$. De igual manera, la SCE se puede expresar en su forma desarrollada para obtener una idea intuitiva sobre la estimación de la forma funcional original:

$$SCE = \Sigma_{i=1}^{n} (y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}))^2 \tag{17}$$

Por ejemplo: para obtener los resultados de la regresión para dos variables independientes mediante el MCO, se procede de la siguiente manera:

- $\hat{y}_1 = b_0 + b_1 x_{1i} + b_2 x_{2i}$

La SCE, resultado de la estimación de $\hat{y_1}$, en el caso de dos variables independientes, $b_1 x_{1i}$ y $b_2 x_{2i}$, se puede expresar de la siguiente manera, tomando en cuenta que el resultado de la relación entre las variables independientes y la variable dependiente observada, y_i:

El desarrollo extenso del MCO es resultado de la aplicación de cálculo diferencial, donde se debe tener en cuenta un sistema de tres ecuaciones lineales y 3 incógnitas, b_0, b_1, b_2, las expresiones resultantes son las siguientes:

$$\left.\begin{aligned} nb_0 + b_1 \Sigma_{i=1}^{n} x_{1i} + b_2 \Sigma_{i=1}^{n} x_{2i} &= \Sigma_{i=1}^{2} y_i \\ b_0 \Sigma_{i=1}^{n} x_{1i} + b_1 \Sigma_{i=1}^{2} x_{1i}^2 + b_2 \Sigma_{i=1}^{2} x_{1i} x_{2i} &= \Sigma_{i=1}^{n} x_{1i} y_i \\ b_0 \Sigma_{i=1}^{2} x_{2i} + b_1 \Sigma_{i=1}^{n} x_{1i} x_{2i} + b_2 \Sigma_{i=1}^{n} x_{2i}^2 &= \Sigma_{i=1}^{n} x_{2i} y_i \end{aligned}\right\} \quad \text{Sistema de Ecuaciones Lineales} \tag{18}$$

Ejercicio en `R`: Utilizando los datos del ejemplo anteriormente mencionado, el comando en `R` para estimar los coeficientes del modelo de regresión lineal es `lm()`. Usandose como ejemplo la siguiente forma:

```
# Usando la función y unas variables para usar
lm(ventas ~ tamaño + tráfico - costos + renta + residentes,
   data = data.frame de los datos)
```

De esta forma, el modelo de regresión lineal múltiple estimado sería a la siguiente ecuación:

- $\hat{ventas}_i = b_0 + b_1(tamaño) + b_2(tráfico) - b_3(costos) + b_4(renta) + b_5(residentes)$

Para almacenar los datos del modelo, a fin de realizar las pruebas pertinentes más adelante, se asigna el nombre a los resultados mismos:

```
# Se guarda en una variable
regresion = lm(ventas ~ tamaño + tráfico - costos +
                 renta + residentes,
               data = data.frame de los datos)
```

La interpretación de los resultados del sistema es la siguiente: en la primera ecuación la variable observada depende de los coeficientes b_1, b_2, asociados a las observaciones de las variables independientes x_1, x_2, y una constante b_0, asociada al número de observaciones n.

En la segunda ecuación, la relación entre la variable independiente y la variable dependiente, x_{1i}, y_i, está explicada por la constante asociada a x_{1i}, las observaciones de x_{1i}, elevadas al cuadrado asociadas a b_1 y el comportamiento entre las dos variables independientes, x_{1i}, x_{2i}, asociadas a b_2.

En la tercera ecuación, la relación entre la variable independiente y la segunda variable dependiente, x_{2i}, y_i, está explicada por la constante asociada a x_{2i}, las observaciones de x_{2i}, elevadas al cuadrado asociadas a b_2 y el comportamiento entre las dos variables independientes, x_1, x_2, asociadas a b_1.

En conclusión, de la representación de la regresión múltiple se infiere, que el coeficiente asociado a la variable explicativa correspondiente, es decir, en el caso de la primera variable independiente, x_{1i}, b_1, está explicada por la misma variable al cuadrado, y en el caso del otro coeficiente, b_2, está explicado por la asociación entre las variables independientes. Lo que se espera, en la regresión es que los dos coeficientes asociados a cada variable dependiente de forma significativa. Lo anterior es resultado, de minimizar los errores asociados a la estimación de la variable independiente en relación a la variable observada.

2.1 ESTIMACIÓN DEL MCO MÚLTIPLE MEDIANTE NOTACIÓN MATRICIAL

La estimación de los coeficientes de las variables independientes mediante el MCO, en su notación matricial, permite visualizar de forma simplificada las operaciones necesarias; esto permite intuir el proceso de estimación de los coeficientes:

- $\hat{y}_1 = b_0 + b_1 x_{1i} + b_2 x_{2i}$

La notación matricial de la expresión anterior es la siguiente: tenemos los dos vectores a estimar la variable dependiente, Y, los coeficientes, β, y los errores de estimación, ε:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \hat{Y} = \begin{bmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_n} \end{bmatrix} \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix} \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{19}$$

Las variables independientes, X, se organizan matricialmente tomando en cuenta su dimensiòn expresada mediante $k - filas$ por $n - columnas$, màs la constante, b_0, representada por una constante numérica igual a 1.

$$\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots x_{k1} \\ 1 & x_{21} & x_{22} & \cdots x_{k2} \\ & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots x_{kn} \end{bmatrix} \tag{20}$$

La construcción de la expresión en su forma matricial reducida es la siguiente:

$$Y = X\beta + U \tag{21}$$

La estimación objetivo del modelo, busca obtener los coeficientes estimados del modelo en relación a las variables independientes, para explicar la variable dependiente, $\hat{Y}$, y su notación es la siguinete:

- $\hat{Y} = X\hat{\beta}$

Donde, la matriz de variables independientes, X, está asociada al vector de coeficientes estimados, $\hat{\beta}$.

La diferencia entre el modelo estimado en su forma matricial y la variable observada, permite obtener los errores derivados de la estimación:

- $Y - \hat{Y} = \varepsilon$

Es decir:

- $\varepsilon = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki})$

Al aplicar el método de MCO, se debe recordar que se minimiza la suma de los errores al cuadrado, SEC:

- $SEC = \Sigma_{i=1}^{n} e_i^2$

Al minimizar respecto al vector de los coeficientes, β, se tiene la siguiente notación matricial reducida:

$$\frac{\partial s}{\partial \beta} = X^T Y - X^T Y + 2(X^T X\beta) \tag{22}$$

$$\frac{\partial s}{\partial \beta} = -2X^T Y + 2(X^T X\beta) = \overrightarrow{0} \tag{23}$$

Para obtener los coeficientes estimados, se despeja β:

- β=$(X^T X)^{-1} - X^T Y$

Entonces, $\hat{\beta}$, es igual a la matriz inversa resultante de la multiplicación entre la matriz transpuesta, X^T, y la matriz, X, menos la matriz transpuesta, X^T, multiplicada por el vector de valores dependientes, Y,. El coeficiente estimado, $\hat{\beta}$, representa el **efecto de un aumento en una unidad de la variable independiente sobre la respuesta de la variable dependiente**, cuando las otras variables independientes se mantienen constantes.

3. LAS PROPIEDADES DE LOS ERRORES

Los estimadores o coeficientes obtenidos tienen propiedades esenciales que permiten una inferencia estadística apropiada, se deduce que la sumatoria de los errores en una serie son igual a cero:

$$\Sigma_{i=1}^{n} e_i x_{ij} = 0 \tag{24}$$

- Donde: $j = 1, 2, 3, \cdots, k$

La covarianza entre los errores y las variables explicativas a medida que aumenta el número de observaciones es igual a cero:

$$Cov = (e_i, x_{ij}) = 0 \tag{25}$$

En el caso del sesgo, se define como la diferencia entre la media del estimador y el verdadero valor del parámetro a estimar. En econometría, se utiliza la varianza residual de los errores, el cual es insesgado al estar entorno a la misma varianza. En este caso se tiene[12]:

$$s_r^2 = \frac{1}{n-(k+1)} \Sigma_{i=1}^{n} e_i^2 \tag{26}$$

La interpetación de los fenómenos económicos mediante un modelo econométrico depende de la robustez de los resultados obtenidos en la estimación. La interpretación inicia con la verificación de la eficiencia de los resultados mediante la inferencia estadística. Cuando se realiza la inferencia en un modelo de regresión múltiple se debe de verificar la estabilidad de los coeficientes y su poder explicativo del modelo.

La distribución de los coeficientes, al igual que en la regresión simple, se distribuyen como una normal, es decir, la media es igual a cero y la desviación estándar es igual a uno.

$$\hat{beta} \sim N(0, 1)$$

Este comportamiento **asegura que los coeficientes estimados sigan una trayectoria normal y no sigan un comportamiento errático que genere problemas en la estimación a medida que aumentan las observaciones**.

[12] En `R`, el comando para obtener el vector de residuales de la estimación es `residuals`. El cual se puede guardar en una variable como acontinuación se muestra: `errores = resultado$residuals`.

El análisis de probabilidad sobre los coeficientes, para identificar la influencia de cada variable parte de la hipótesis planteada desde el diseño del modelo y su forma funcional. El contraste de hipótesis, se construye mediante una t de *Student*, con k grados de libertad, la prueba muestra las siguientes posibilidades:

- Hipótesis nula: $H_0 : \beta_i = O$
- Hipótesis alternativa: $H_1 : \beta_i \neq 0$

Al aplicar el contraste de hipótesis, cuando la probabilidad de cometer el **error tipo A** es elevada, es decir, rechazar la hipótesis nula, H_0, cuando es verdadera y aceptar la hipótesis alternativa, H_1, cuando esta última es falsa, entonces, lo correcto es aceptar la hipótesis nula, H_0; de ahí se puede inferir que la variable independiente, X_i, asociada a su coeficiente tiene un efecto nulo, es decir, no influye sobre la variable dependiente.

El diseño de la prueba es el siguiente, la distribución del valor de los coeficientes cuando se acepta la hipótesis nula, H_0, se distribuyen de la siguiente forma: para $n > 30$ observaciones la distribución es $t_{n-(k+1)}$, bajo una probabilidad del 95 % se encuentra en el intervalo $[-2, 2]$ y entonces se acepta la hipótesis nula. Si $t > 2$, entonces se rechaza la hipótesis nula, H_0, y se puede inferir estadísticamente que las variables independientes influyen en la variable dependiente, es decir, se acepta la hipótesis alternativa, H_1. El contraste de hipótesis señala que la probabilidad de cometer el **error tipo A** es nulo, por lo tanto, se puede rechazar la hipótesis nula, H_0, y aceptar la hipótesis alternativa, H_1.

- $H_1 : \beta_i \neq 0$

El criterio del intervalo de confianza está diseñado de la siguiente forma:

$$P(\hat{\beta}_i - t\frac{\alpha}{2}SE(\hat{\beta}_i) \leq \beta_i \geq \hat{\beta}_i + t\frac{\alpha}{2}SE(\hat{\beta}_i)) = 1 - \alpha \tag{27}$$

El criterio muestra la probabilidad de que el verdadero β_i se encuentra en el intervalo entre el coeficiente estimado, $\hat{\beta}_i$, y 2 desviaciones estándar, SE, a la derecha y a la izquierda. Cuando se tiene un intervalo de confianza de $\alpha = 0.05$, se plantea que hay un 95 % de confianza de que el valor verdadero, para cada coeficiente, se encuentre dentro del área de aceptación[13].

La matriz de varianzas-covarianzas de los coeficientes en su forma matricial reducida es la siguiente:

$$Cov(\hat{\beta}_i) = \sigma^2(X^T X)^{-1} \tag{28}$$

De la función anterior es necesaria la estimación de la varianza, σ^2, en la estimación del modelo, se espera que la varianza de los residuos sea el valor verdadero de la varianza de los estimadores, es decir, que la varianza de las variables incluidas en el modelo explique los errores de la estimación:

[13]Retomando el ejemplo en `R`, el comando necesario para obtener los estadísticos, tales como la probabilidad de los coeficientes del modelo, es el siguiente: `summary(resultado)`.

- $E(\hat{S}_e^2) = \sigma^2$

Este resultado permite establecer que la elección de las variables en la estimación del modelo, sea la especificación correcta. Ya que, explica las desviaciones de la variable dependiente respecto a la estimada[14].

Una forma de medir el poder explicativo del modelo es el contraste F, muestra si las variables explicativas en conjunto explican las variaciones de la variable independiente. Se ha demostrado que los coeficientes $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$, y además, siguen una distribución F de $Fisher$ dado la siguiente forma:

$$\frac{\frac{\Sigma_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{k}}{\frac{\Sigma_{i=1}^{n} e_i^2}{n-(k+1)}} \sim F_{k,n-(k+1)} \tag{29}$$

El resultado muestra la proporción en que la varianza de los coeficientes explica la variación en los errores; cuando se acepta la hipótesis nula, se debe a dos factores:

1. Las variables no influyen en la variable independiente.
2. Existe dependencia no lineal entre la variable explicada y algún regresor.

Cuando se rechaza la hipótesis nula, en el contraste de la prueba F, muestra que la variable dependiente está explicada por alguna de las variables independientes. Para conocer de forma específica las variables con poder explicativo relativo a las otras variables, es necesario revisar los contrastes individuales mediante la t de *student*.

En la aplicación de los contrastes de la prueba F de $Fisher$ se presentan los siguientes casos:

1. Cuando el contraste F es significativo y todos los coeficientes individuales de acuerdo al contraste de la prueba t de *student* también son significativos, en este caso, **todas las variables independientes son significativas para explicar el comportamiento de la variable dependiente**.
2. Si el contraste F es significativo y sólo algunos de los coeficientes individuales son significativos de acuerdo al contraste de la t de *student*, **las variables no significativas deben ser eliminadas del modelo**. Otra solución es realizar una transformación y estimar nuevamente para verificar si la relación entre variables no es lineal.
3. Cuando el contraste F es significativo y por el otro lado, cuando ninguno de los coeficientes asociados a las variables es significativo de acuerdo al contraste t de *student*, entonces **podría estar presente un problema de multicolinealidad**. Esta última es el resultado de una correlación alta entre las variables independientes; entonces, la especificación del modelo requiere una elección eficiente de las variables.

[14] En el ejemplo en `R`, la matriz de varianzas-covarianzas se obtiene de la siguiente función: `vcov(resultado)`.

En la tabla **ANOVA**, se puede evaluar los resultados mediante el **Test F**[15]:

$$\frac{\hat{S}_e^2}{\hat{S}_r^2} \tag{30}$$

El **Test F** muestra la proporción en el que la varianza de los errores determina el poder explicativo del modelo. La notación matricial de la prueba, muestra que la diagonal de la matriz conocida, arroja los valores de la varianza, σ^2:

$$D(X^T X)^{-1} \rightarrow \begin{bmatrix} d_{00} & & & \\ & d_{11} & & \\ & & d_{ii} & \\ & & & d_{kk} \end{bmatrix} \tag{31}$$

De esta forma, la distribución de los coeficientes estimados es la siguiente:

- $\hat{\beta}_i \sim N(\beta_i \sigma \sqrt{d_{ii}})$

En donde, la desviación de los coeficientes tienen una distribución normal:

- $\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{d_{ii}}} \rightarrow N(0, 1)$

La desviación entre el coeficiente estimado, $\hat{\beta}$, y el coeficiente, β_i, en proporción a la integración en diagonal conocida se comporta como una normal[16].

[15] El comando para obtener la tabla ANOVA del ejemplo es através de la siguiente función: `anova(resultado)`.

[16] La prueba F en `R` se realiza con la siguiente función `var.test(resultado)`.

4. PRUEBAS DE DIAGNÓSTICO

La información relevante en los modelos de regresión múltiple, está contenido en las variables seleccionadas. Los modelos operan bajo el supuesto de que el modelo contiene todas las variables relevantes para explicar el modelo. En este sentido la realización de pruebas de diagnóstico sobre la sección eficiente de las variables incluidas en el modelo es necesario. La omisión de las variables relevantes en el modelo, es un problema relevante en la especificación del modelo y en este sentido se puede generar problemas de multicolinealidad.

Al iniciar el capítulo se planteó que el primer paso es la especificación del modelo, la sección de las variables para la conformación del modelo, se realiza con los referentes que ofrece la teoría económica. Como se ha señalado, las variables referentes en estos modelos no especifican como podrían conformar un modelo econométrico. El primer paso, es revisar la teoría para contrastar las variables relevantes que explican el objeto de estudio desde esa perspectiva. El siguiente paso es realizar una prueba de omisión de variables, supongamos que la teoría señala que la regresión correcta incluye dos variables.

- $Y = X_1\beta_1 + X_2\beta_2 + U$

Finalmente, tras un proceso de elección, el modelo estimado es:

- $Y = X_1\beta_1 + U$

El siguiente paso es plantear la hipótesis nula de la omosión de variables:

- $H_0 : \beta_2 = 0$

Posteriormente se realiza una prueba de contraste F para estimar el poder explicativo del modelo, en un caso se estimará la prueba al modelo estimado y una prueba para el modelo que incluye la variable omitida. El rechazo de la hipótesis nula en este caso mostrará que fue omitida una variable relevante.

De igual manera, cuando se incluyen variables irrelevantes en el modelo, es necesario realizar pruebas para la especificación del modelo. De hecho, cuando se aplica una metodología donde se parte de la especificación más general se realizan estas pruebas para llegar a un modelo más específico.

La prueba de inclusión de variables irrelevantes consiste en probar la hipótesis:

- $H_0 : \beta_2 = 0$