ECONOMETRÍA
EN TU IDIOMA

HUGO BRISEÑO • DOLORES LUQUÍN

CIENCIAS ECONÓMICAS Y EMPRESARIALES

UNIVERSIDAD Panamericana

ECONOMETRÍA
EN TU IDIOMA

ECONOMETRÍA
EN TU IDIOMA

HUGO BRISEÑO • DOLORES LUQUÍN

CIENCIAS ECONÓMICAS Y EMPRESARIALES

UNIVERSIDAD
Panamericana

Primera edición, 2025
Edición con nueva cubierta, 2026

Versión electrónica

Título: Econometría en tu idioma
Autor: Hugo Briseño, Dolores Luquín
Facultad de Ciencias Económicas y Empresariales

Responsable editorial: Manuel Bernal Coronel
Diseño de portada: Ivonne Lara Alba
Cuidado editorial: Santi Ediciones

ISBN: 978-607-8826-91-9

Universidad Panamericana, Campus México
Jerez 10, Insurgentes Mixcoac, Benito Juárez,
Ciudad de México, México, C.P. 03920
Conmutador: +52 55 5482 1600
www.up.edu.mx

Impreso en México / Printed in Mexico.

Para el Dr. Antonio Ruiz Porras,
gracias por el conocimiento transmitido
a tantas generaciones.

CONTENIDO

PRÓLOGO

En 2005 un grupo de estudiantes de los últimos semestres de Finanzas me buscaron y me pidieron que les impartiera un curso de econometría. Como profesor recién llegado a la UP, acepté. En aquel entonces, la econometría despertaba más curiosidad que certeza entre los futuros financieros: una mezcla entre economía, matemáticas y estadística que prometía convertir los números en historias reales sobre mercados e inversiones.

Entendí el interés de los alumnos, ya que la materia no aparecía en su plan de estudios, pero a los alumnos les preocupaba terminar sus estudios y no tener conocimientos de econometría.

El curso era informal, sin compromiso de los alumnos por asistir, ya que no había una asistencia o calificación que reportar. Aun así, el entonces director de la carrera, Javier Castañeda, nos apoyó con la gestión de un salón para impartir la clase.

El curso fue un desastre. Solamente tuve dos alumnos, pero entre esos dos únicos asistentes se encontraban el doctor Hugo Briseño, autor del presente libro.

Aquel curso informal, quiero pensar, sembró una inquietud que con el tiempo se transformó en pasión y rigor académico.

Desde entonces, la econometría ha dejado de ser una asignatura informal y se ha convertido en una herramienta para los profesionales de las finanzas. Hoy, que la carrera de Administración y Finanzas ha crecido, creció también el interés por entender, aplicar e interpretar modelos econométricos. Ya no se trata solo de "hacer regresiones", sino de comprender qué nos dicen los datos sobre los fenómenos económicos y financieros que moldean nuestro entorno.

El libro que el lector tiene ahora entre sus manos, *Econometría en tu idioma*, representa justo ese espíritu: el de acercar una disciplina que es compleja, envuelta en fórmulas y símbolos, al lenguaje claro, cotidiano y significativo para los estudiantes. Es un

texto que combina el rigor metodológico con ejemplos prácticos, que enseña tanto a estimar como a interpretar, y que busca hacer de la econometría una aliada —no un obstáculo— para quienes se preparan para analizar mercados, evaluar riesgos o tomar decisiones financieras.

Este trabajo es también una invitación para las nuevas generaciones de financieros y administradores a no temerle a los modelos econométricos o a la programación, sino a ver en ellos la posibilidad de comprender mejor la realidad económica que los rodea.

Celebro la publicación de este libro y la madurez intelectual que representa para nuestra comunidad académica. Que sirva no solo como guía técnica, sino como punto de partida para seguir cultivando el análisis riguroso y la pasión por descubrir, en los datos, las historias que nos cuentan los mercados.

Dr. Israel Macías López
Facultad de Ciencias Económicas y Empresariales
Universidad Panamericana
Guadalajara, México

SOBRE LOS AUTORES

Hugo Briseño Ramírez
Universidad Panamericana
https://orcid.org/0000-0001-8465-8683.
Correo: hbriseno@up.edu.mx

Hugo Briseño es doctor en Ciencias Económico Administrativas con orientación en Políticas Públicas por la Universidad de Guadalajara; maestro en Ciencias Sociales por El Colegio de Sonora; maestro en Valuación por la Universidad de Guadalajara; licenciado en Administración y Finanzas por la Universidad Panamericana; y especialista en Fintech por el Instituto de Estudios Bursátiles (IEB) de Madrid. Miembro nivel 1 del Sistema Nacional de Investigadoras e Investigadores (SNII) contando con varias publicaciones científicas indexadas en la base de datos Scopus. Autor del libro *Indicadores Financieros Fácilmente Explicados*. Aprobó el examen nivel 1 del CFA Institute y está certificado por la empresa de sensores biométricos iMotions en investigación del comportamiento humano.

Es director del Observatorio de Ciudades Hidroadaptativas de la Universidad Panamericana. Se desempeñó como secretario de investigación de la Facultad de Ciencias Económicas y Empresariales del campus Guadalajara. Fue jefe de la Academia de Finanzas en la misma institución. Trabajó como coordinador en la Secretaría de Planeación, Administración y Finanzas del Gobierno del Estado de Jalisco. En la Universidad Panamericana imparte materias de corte financiero y económico a nivel de licenciatura, maestría y doctorado. Sus intereses de investigación están centrados en métodos cuantitativos aplicados al análisis de asuntos públicos, gestión del

agua para uso urbano y valuación de activos. Realiza actividades de consultoría a empresas. Es miembro del Consejo Académico del Agua de Jalisco, del Consejo de Cuenca del Río Santiago y ha participado como consejero en la Comisión Tarifaria del SIAPA así como en diferentes organismos gubernamentales y en el Congreso del Estado en temas relacionados con el agua. Ha sido vicepresidente del Consejo Directivo del IMEF grupo Guadalajara y vicepresidente del Consejo Editorial de la revista News IMEF. Es presidente del IMEF grupo Guadalajara para 2025.

Dolores Luquín-García
Universidad Panamericana
https://orcid.org/0000-0003-1976-7378
Correo: dluquin@up.edu.mx

Dolores Luquín es doctora en Estudios Económicos por la Universidad de Guadalajara, con maestría en Ingeniería Industrial con énfasis en Optimización, licenciada en Administración y Mercadotecnia. En 2025 terminó la Especialidad en Ciencia de Datos en la UP. Desde 2013 es parte del claustro académico de la Universidad Panamericana, ha impartido clases tanto la Academia de Mercadotecnia como en la de Matemáticas.

Fue la investigadora responsable de dos proyectos financiados por el fondo de Fomento a la investigación UP 2014 y UP 2016 y es parte de un equipo en 2025. Coordinó el desarrollo del software "Geo-Spatial Economy", el cual obtuvo mención honorífica en el Galardón "Manuel López Cotilla", otorgado por la institución Jalisco Tecnológico (Jaltec). Coodinadora asimismo de otros dos softwares para el análisis espacial y del libro *Problemario de investigación de Operaciones: con aplicación a las ciencias económicas-administrativas*, publicado en físico y en electrónico. Actualmente, es la líder del Business Research Center (BRC) de la Universidad Panamericana Campus Guadalajara, encargándose de cinco laboratorios para la investigación económico-administrativa: Innovation Lab, Retail Lab, Neuro Lab, Behavioral Lab y Finance Lab. Sus áreas de interés son: Factores económicos e interdisciplinares de localización de empresas, Geomercadotecnia, Modelado matemático.

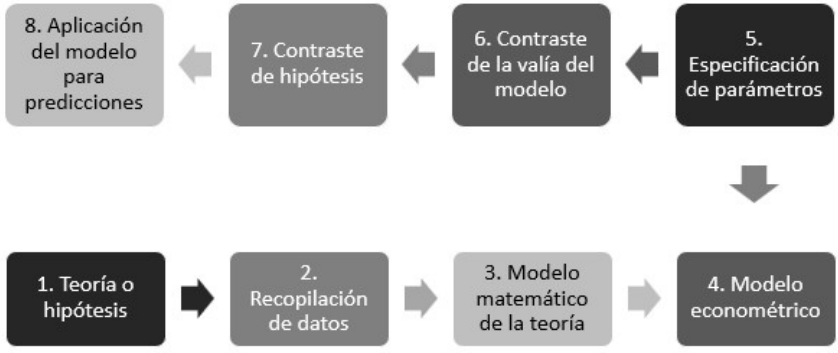
MATERIAL DIDÁCTICO

Escanea el siguiente QR para encontrar ejemplos prácticos realizados desde el software estadístico y econométrico Gretl. En estos encontrarás desde la manera en la que se recomienda estructurar una base de datos hasta la realización de modelos de mínimos cuadrados ordinarios, panel de datos, probit y logit. Asimismo, podrás encontrar algunas notas técnicas que te ayudarán a observar de manera resumida algunos capítulos del libro.

BASES DE DATOS Y TIPOS DE VARIABLES EN MODELOS ECONOMÉTRICOS

La econometría es la ciencia social donde se aplican la teoría económica, la estadística inferencial y la economía matemática para el análisis de los fenómenos económicos y financieros (Gujarati y Porter, 2010b, p. 3). Para los estudiantes de la Facultad de Ciencias Económicas y Empresariales la relevancia de estudiar econometría estriba en que, como parte de su empleo, deberán pronosticar ventas, tasas de interés, gráficas de acciones bursátiles o estimar las funciones de oferta y demanda para encontrar equilibrios de mercado, así como pronosticar precios y calcular diferentes elasticidades. La metodología que se sigue para la formación de modelos econométricos se describe a continuación:

Figura 1. Proceso de formación de modelos econométricos



Fuente: elaboración propia con base en Gujarati y Porter (2010b, p. 3).

También se deben considerar los supuestos que deben cumplir los modelos econométricos, los cuales se especifican en los siguientes puntos:

1. Los modelos econométricos indican una relación de causalidad entre las variables.
2. Son consistentes en el tiempo.
3. Se debe considerar que existe el error estocástico, es decir, el error que surge de los efectos no capturados por las variables económicas.
4. En un modelo econométrico se deben establecer los signos de cada término, los cuales deben ser consistentes con la teoría económica.
5. La modelación es con base en estimados, es decir, con información incompleta.
6. Se parte de que siempre habrá diferencias entre lo real y lo estimado, sin importar que tan elaborado sea el modelo.

Lo que podríamos preguntarnos ahora es ¿qué es un modelo? Una definición ampliamente aceptada es una representación simplificada de la realidad. Sin embargo, esta definición no expresa a cabalidad el tipo de modelos que se buscan construir. Entonces, un modelo económico se puede definir, siguiendo a Hernández (1995), como la expresión matemática de una teoría económica, ya que se requiere un lenguaje matemático para especificarlo. Un modelo econométrico va más allá, puesto que se trata de un modelo económico que conjunta la estadística, las matemáticas y la teoría económica, factores necesarios para su aplicación empírica.

Tipos de variables: cualitativas, cuantitativas, discretas y continuas

Las variables se refieren a los insumos controlables del modelo, es decir, las alternativas de decisión especificados por quién construye el modelo. La nomenclatura de las variables más comúnmente utilizadas en econometría es x e y. Es importante resaltar que la relación causal entre x e y, si es que existe, debe estar basada en la teoría económica pertinente.

Dichas variables pueden medirse en cuatro escalas distintas:

Escala de razón: la mayor parte de las variables económicas pertenecen a esta categoría. Deben poder expresarse como una división (razón) entre dos variables, como una resta entre dos variables y también deberán poder ser ordenadas, ya sea de mayor a menor o de menor a mayor. En este caso, se tratan generalmente de variables cuantitativas, o medibles, como puede ser el PIB, las ventas, los gastos o un precio.

Escala de intervalo: en este caso, no se cumple con el primer supuesto de la escala de razón, pero sí cumplen con las otras dos: se puede expresar como una resta o se pueden ordenar. Un ejemplo claro de este tipo de variables son los años, ya que se puede definir el intervalo que ha pasado entre dos años o se pueden ordenar los años, pero no tiene sentido expresarlos como una división o razón.

Escala ordinal: las variables de este tipo satisfacen el supuesto de que se pueden ordenar, pero no se pueden expresar como razón (o división) ni se pueden restar. Los lugares de una carrera de F1 son un ejemplo de variables ordinales, así como el grado de estudios (primaria, secundaria, preparatoria).

Escala nominal: las variables de esta categoría no cumplen con ninguno de los supuestos de la escala de razón, es decir, no se pueden expresar como una división, ni se pueden restar u ordenar. Se les denominan también variables cualitativas o *dummy*. Ejemplos de estas variables pueden ser el género, la religión, el país de origen o el estado civil. Para su uso en modelación matemática, deben transformarse a valores numéricos, asignados de acuerdo con el número de categorías de cada variable, pero sin que este número represente una cantidad u orden específicos. Existen varios métodos para llevar a cabo dicha transformación: escalamiento, estandarización, normalización, transformación Box-Cox, transformación Yeo-Johnson o binarización.

Ahondando en la llamada variable *dummy*, también pueden recoger efectos temporales o espaciales dentro de las ecuaciones. Por lo general, asumen únicamente valores discretos, es decir, que se pueden contar (1,2,3...,n) pero no asumen valores decimales. Un tipo especial de las variables *dummy*[1] son las variables binarias, que solo pueden tomar valores de 0 y 1.

En caso de que el número de categorías sea múltiple, se definen $m - 1$ variables ficticias[2], y en cada variable se establece como 1 solo la categoría que se quiere definir, siendo el resto representadas con ceros. Por ejemplo, si se establece una variable *dummy* de grado de educación máxima, las *dummy* quedarían definidas como sigue:

Figura 2. Ejemplo de variable *dummy*

		D_1	D_2	D_3	D_4	D_5
Nivel de estudios	Primaria	1	0	0	0	0
	Secundaria	0	1	0	0	0
	Preparatoria	0	0	1	0	0
	Universidad	0	0	0	1	0
	Posgrado					

Fuente: elaboración propia.

Las variables, asimismo, pueden tener otra clasificación:

Endógenas: son aquellas que son explicadas dentro del modelo y pueden ser explicadas y también influir en otras variables del modelo.

Exógenas: son aquellas variables que se determinan fuera del modelo, se incluyen para explicar las endógenas. Influyen en las variables del modelo, pero nunca son influidas por otras del modelo.

1 Se utiliza el término *dummy* en singular aun cuando se habla del término variables en plural.

2 La definición de $m - 1$ variables es para evitar la llamada **trampa de la *dummy***, en la que existe multicolinealidad perfecta, o en palabras simples, todas las variables se relacionarían entre sí y no sería posible realizar una regresión significativa.

Tipos de datos

Para llevar a cabo las regresiones se requieren datos. Los datos son valores observados o registrados de una variable, que puede ser económica. Existen tres tipos de datos que se encuentran disponibles para el análisis:

Datos de series de tiempo: se trata de un conjunto de observaciones que toma una variable en cortes de tiempo distinto. Pueden tratarse de los precios de las acciones, el Índice Nacional de Precios al Consumidor (INPC), el producto interno bruto (PIB) o el Censo de Población y Vivienda. Las temporalidades con la que se recolectan los datos pueden ser diaria, semana, trimestral, anual, quinquenal o cualquier otro corte temporal en que se generan las observaciones. Se espera de las series de tiempo que sean estacionarias, lo que significa que su media (promedio) y varianza no varíen sistemáticamente a través del tiempo. Las observaciones de series de tiempo se indican con el subíndice *t*, por ejemplo Y_t, X_t.

Datos de corte transversal: son datos de una o más variables que se recolectan en el mismo punto de tiempo, como pueden ser los resultados de una encuesta de investigación recolectada por varias universidades en un mismo mes, la temperatura de un día específico en distintas ciudades o como en la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) los ingresos y gastos realizados por los mexicanos cada año par. Los datos de corte transversal se caracterizan por el subíndice *i*, como en Y_i, X_i.

Datos de panel o longitudinales: este tipo de datos combina características de los datos de corte transversal y los de series de tiempo. Por ejemplo, en el censo económico que se realiza de forma quinquenal, se puede observar el aspecto de serie de tiempo ya que sus datos se generan en períodos quinque-

nales, pero en los aspectos de corte transversal, como puede ser por tipo de industria o por municipio. Las observaciones de los datos de panel se denotan con doble subíndice: Y_{it},X_{it}.

Diferencias entre variable, parámetro, estimador y dato

Un parámetro es un coeficiente matemático que acompaña a las variables del modelo; son magnitudes consideradas constantes dado un fenómeno específico y generalmente se les denomina parámetros estructurales. En este libro, se especificarán como betas (β) y se irán distinguiendo con un subíndice según vaya aumentando el modelo, por ejemplo β_0 corresponde al término constante, β_1 al coeficiente de la variable 1, β_2 al coeficiente de la variable 2 y así consecutivamente. En econometría, se suele exigir un comportamiento matemático, que como vimos al inicio de este capítulo, puede ir desde tomar un cierto signo o estar restringidos a ciertos valores, como 0 o 1, mayor o igual a 1, etcétera. Cuando hablamos de parámetro nos referimos a la medida de sensibilidad que queremos conocer, pero nos referimos a estimador al dato que nos permite acercarnos a ese valor.

Una variable, como ya habíamos considerado con anterioridad, corresponde a información que pueden tomar un valor determinado. En los modelos econométricos, les corresponden las nomenclaturas de *x* e *y*. Por lo común, las variables *x* son independientes, de preferencia se debe buscar que no se correlacionen entre ellas para que el modelo econométrico sea especificado de forma correcta. En el presente libro, se especificarán las con subíndices de ser necesario: x_1, x_2, ..., x_n. La variable dependiente generalmente se denomina como *y*.

En cambio, los datos representan la información que se registra para integrar las observaciones contenidas en las variables. Es importante asegurar que la fuente de los datos es confiable, profusa y de calidad (Gujarati y Porter, 2010b, p. 5).

Estructura básica de una base de datos econométrica

A continuación, se presentará un ejemplo de una base de datos que se puede formar para llevar a cabo un modelado econométrico:

Base de datos de series de tiempo: en este caso, se presenta la información de la cantidad de unidades económicas (empresas) presentes en el país en los distintos censos económicos de 2003, 2008, 2013 y 2018.

Tabla 1. Cantidad de empresas de los censos económicos de 2003, 2008, 2013 y 2018

Año censal	Entidad	Actividad económica	UE Unidades económicas
2003	00 Total Nacional	Total nacional	3005157
2008	00 Total Nacional	Total nacional	3724019
2013	00 Total Nacional	Total nacional	4230745
2018	00 Total Nacional	Total nacional	4800157

Fuente: elaboración propia.

Base de datos de corte transversal: el ejemplo presentado es también la cantidad de empresas de los censos económicos, pero en este caso se consideró solo el de 2018 y más bien se especificó la información por estado de la República mexicana:

Tabla 2. Cantidad de empresas de los censos económicos por estado (2018)

Año censal	Entidad	Actividad económica	UE Unidades económicas
2018	01 Aguascalientes	Total estatal	53939
2018	02 Baja California	Total estatal	105215
2018	03 Baja California Sur	Total estatal	30601
2018	04 Campeche	Total estatal	35275
2018	05 Coahuila de Zaragoza	Total estatal	95230
2018	06 Colima	Total estatal	33566
2018	07 Chiapas	Total estatal	186996
2018	08 Chihuahua	Total estatal	106430
2018	09 Ciudad de México	Total estatal	427959
2018	10 Durango	Total estatal	56236

Fuente: elaboración propia.

Base de datos de panel: continuando con el mismo ejemplo de los censos económicos, se generó la cantidad de empresas de los censos de 2003 a 2018 para cada Estado de la República mexicana:

Tabla 3. Cantidad de empresas de los censos económicos por estado (2003 a 2018)

Año censal	Entidad	Actividad económica	UE Unidades económicas
2018	01 Aguascalientes	Total estatal	53939
2013	01 Aguascalientes	Total estatal	47449
2008	01 Aguascalientes	Total estatal	40988
2003	01 Aguascalientes	Total estatal	33630
2018	02 Baja California	Total estatal	105215
2013	02 Baja California	Total estatal	95882
2008	02 Baja California	Total estatal	80380
2003	02 Baja California	Total estatal	61812
2018	03 Baja California Sur	Total estatal	30601
2013	03 Baja California Sur	Total estatal	28114

Fuente: elaboración propia.

REGRESIÓN LINEAL SIMPLE

Planteamiento del modelo

Cuando hablamos de regresión asumimos que estamos buscando una línea establecida como $y = mx + b$, en la que tenemos una variable independiente (x), una dependiente (y) y queremos determinar la ecuación de la recta que represente de mejor manera la información disponible.

Llevando a cabo un paralelismo con la ecuación de la recta inicial presentada, $y = mx + b$, esta tiene una de las formas funcionales básicas denominada *nivel-nivel*, la cual se verá con mayor profundidad en el capítulo tres, correspondiente a formas funcionales. En el caso de presentarse una sola ecuación, lineal, estática y completa, se presume una forma funcional simple, lo cual hace alusión a que se incorpora una única variable independiente o explicativa. La forma del modelo lineal simple en econometría se presenta a continuación:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

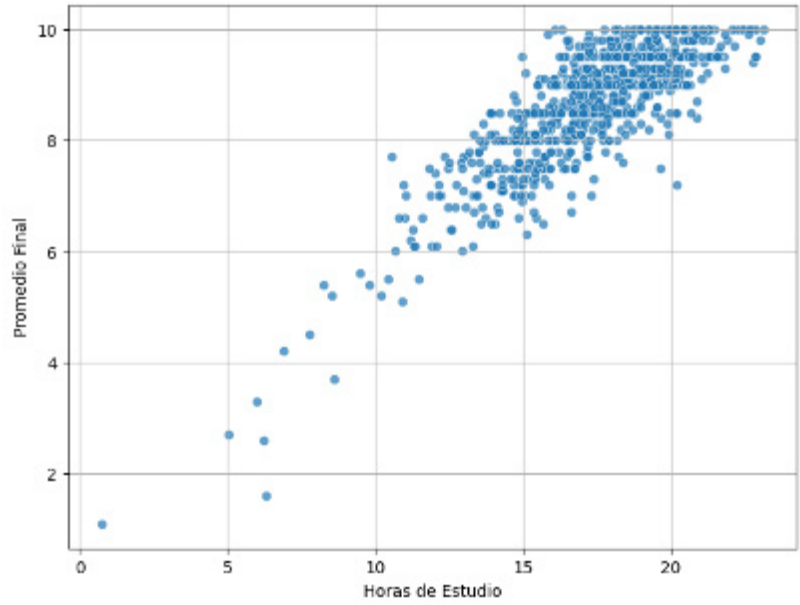
Donde:
y = variable dependiente.
x = única variable independiente.
β_0 = intercepto.
β_1 = pendiente.
ε = término de error aleatorio.

Ejemplo de un modelo lineal simple

Vamos a establecer un ejemplo hipotético fácilmente entendible, con una situación cotidiana. Imaginemos que un profesor universitario quiere descubrir cuál sería la ecuación que describe la relación entre las horas de estudio con el promedio final de las materias que imparte. Elabora un archivo en el que registra información de la materia, año, ID, carrera, horas de estudio y calificación final de la materia.

La gráfica mostrando el diagrama de dispersión de la relación entre las horas de estudio y el promedio final se muestra a continuación:

Gráfica 1. Promedio final vs horas de estudio



Fuente: elaboración propia.

Se establece un modelo lineal simple:

$$\hat{y} = \hat{\beta}_0 + \beta_1 x$$

Donde:
$\hat{y}$ = variable dependiente, las horas de estudio.
x = única variable independiente, el promedio final en nuestro ejemplo.
β_0 = intercepto, valor esperado de promedio final cuando las horas de estudio son cero.
β_1 = pendiente, representa cuánto se espera que aumente el promedio final por cada hora de estudio adicional.

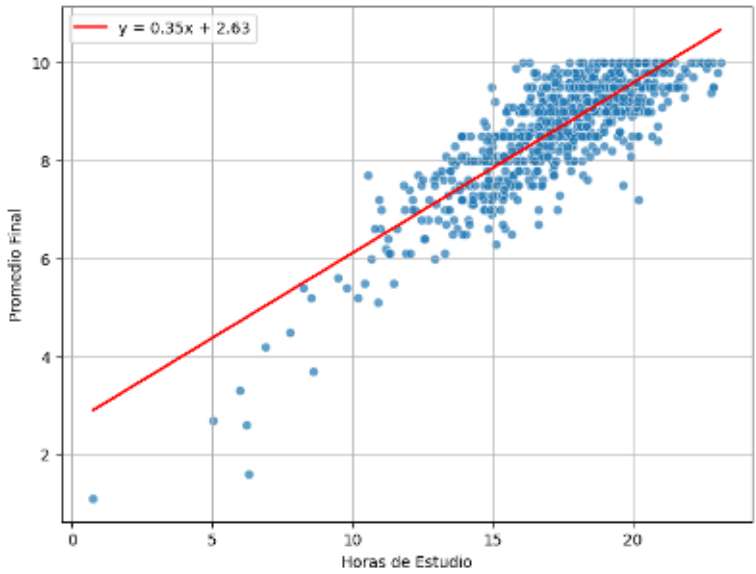
En el ejemplo, el término de error aleatorio no aparece explícitamente porque se estima como el residuo entre la predicción y la verdadera calificación de examen final, es decir, $\varepsilon = y - \hat{y}$

Al realizar el cálculo de la ecuación lineal simple se obtiene que:

$$\hat{y} = 2.63 + 0.35x$$

Lo cual significa que cuando el estudiante no estudia, el valor esperado de promedio final es de 2.63 Asimismo, cada hora de estudio adicional genera un aumento de 0.35 décimas. La gráfica quedaría de la siguiente manera:

Gráfica 2. Promedio final vs horas de estudio



Fuente: elaboración propia.

Pero, ¿cómo se calcularon β_0 y β_1? Además, ¿cómo podemos saber qué tan exactamente mide la ecuación el fenómeno de la relación entre el promedio final con respecto la cantidad de horas de estudio? En la siguiente sección se responden a estas preguntas y además se establecen estimadores adicionales para identificar qué tan significativo es cada uno de los coeficientes calculados en la regresión.

Cálculo de coeficientes

β_1

Se inicia con el cálculo de β_1, ya que este valor será importante para encontrar el valor de β_0. Este estimador identifica la pendiente, o el cambio que ocurre en *y* cuando cambia *x*. Es la misma que en

álgebra identificábamos con la letra *m*. En el ejemplo de las calificaciones era cuánto cambiaba la calificación de promedio final por hora adicional de estudio. La fórmula para calcular este estimador es la siguiente:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Cálculo del numerador: para cada uno de los valores de *x* y de *y* se calcularán dos estimadores, el denominado promedio de *x* ($\bar{x}$) y el promedio de *y* ($\bar{y}$). Se mantendrá el orden de pares ordenados (*x, y*). A cada uno de los valores de *x* se le restará el valor de $\bar{x}$ y cada uno de los valores de *y* se les restará el valor correspondiente de $\bar{y}$. Los valores resultantes de las restas se multiplicarán entre sí. Una vez llevado a cabo el cálculo de cada resta y cada multiplicación, se realizará la suma de todas las multiplicaciones de las diferencias, lo que representa

$\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$.

Cálculo del denominador: las diferencias (restas) de cada uno de los valores de *x* menos su promedio ($\bar{x}$) se elevarán al cuadrado. Una vez realizada la potencia al cuadrado, se sumarán todos los números, esto se representa como

$\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Finalmente, para encontrar $\hat{\beta}_1$ se dividirá el numerador $\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$ entre el denominador $\sum_{i=1}^{n}(x_i - \bar{x})^2$ dando como resultado un único número, que es el estimador que estamos buscando.

Una forma más sintética de calcular el coeficiente $\hat{\beta}_1$ de la regresión es a través de la siguiente fórmula:

$$\hat{\beta}_1 = \frac{Cov(x,y)}{s_x^2} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{n-1}}{s_x^2}$$

Donde:

$\hat{\beta}_1$ = estimador de la β_1.

$Cov(x,y)$ = covarianza entre las variables *x* y *y*.

s_x^2 = varianza muestral de *x*.

β_0

Como se estableció anteriormente, este estimador identifica el valor que la variable dependiente tendría, si el valor de la variable independiente es cero. Es decir, (0,*b*) en el que *b* es el valor de *y* cuando *x* es cero (*x*=0). En nuestro ejemplo, era el valor del promedio final cuando el alumno no tuvo horas de estudio. Se calcula de la siguiente manera:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Donde:
$\bar{y}$ = el promedio de los valores de *y*.
$\bar{x}$ = el promedio de los valores de *x*.

Su cómputo es muy directo una vez establecido el valor de β_1, ya que cuando se calculó dicho parámetro, también se estableció el promedio de *x* ($\bar{x}$) y el promedio de *y* ($\bar{y}$). Se sustituyen los valores de β_1, $\bar{x}$ y $\bar{y}$ para encontrar $\hat{\beta}_0$.

Coeficiente de determinación: R^2 y R^2 ajustado

R^2

Es un indicador que establece en qué porcentaje la variabilidad está explicada por el modelo. Si el R^2 es mayor, tu modelo explicará mejor los datos. La fórmula para calcularlo es:

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Donde SCE es la suma al cuadrado de los términos del error y el SCT es la suma de cuadrados totales. Se calculan:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SCT = SCE + SCR$$

Individualmente,

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Donde:
$\bar{y}$ = promedio de los valores de *y*.
y_i = cada uno de los valores originales de *y*.
$\hat{y}_i$ = el valor estimado de cada uno de los valores de *y*.

SCT es la varianza total del modelo. El SCE explica cuánta varianza queda después de ajustar el modelo lineal, es decir, la diferencia entre el valor predicho por el modelo y el valor real. El SCR mide la diferencia al cuadrado entre los valores predichos y el promedio de los valores reales. El R^2 indica qué tanto de la varianza se explica por el modelo lineal. El R^2 tiene un rango entre 0 y 1, siendo que entre más cercano al 1 los datos se explican completamente con el modelo establecido en la regresión y uno con bajo R^2 establece que el modelo no explica los datos de forma adecuada.

Sin embargo, el problema con el R^2 es que según aumentan el número de variables explicativas ($x's$) el R^2 aumenta, aún si no aportan al modelo. Para evitar esta situación, se lleva a cabo un ajuste, denominado *R^2 ajustado*, el cual veremos a continuación.

R^2 ajustado

Para que el número de variables explicativas no influyan directamente en que aumente el R^2, se calcula el ajustado, que incluye en

su cálculo la cantidad de variables utilizadas en el modelo. Su fórmula se presenta a continuación:

$$R^2_{ajustado} = \left[1 - \frac{(n-1)}{(n-k)}\right](1 - R^2)$$

Donde:
n = cantidad de observaciones. Por ejemplo, el número de filas de la base de datos en Excel.
k = número de betas utilizadas para la predicción.
$n - k$ = se denominan grados de libertad.

Al tener un modelo con R^2 con poco ajuste, dicho modelo no ayuda en la explicación de la varianza de la regresión. Esto es debido a la naturaleza de los fenómenos humanos que se estudian en estas ciencias. Sin embargo, no debemos obsesionarnos con la R^2, depende mucho de la distribución de los datos. En ocasiones no se comportan linealmente y presentan R^2 muy bajas, aunque no por eso dejen de ser resultados relevantes.

Varianza y error estándar de la regresión lineal simple

Partimos del supuesto de que las variables que utilizamos en la regresión lineal simple son variables aleatorias, ya que sus valores cambian si tomamos muestras distintas del universo. Para identificar la variabilidad de las variables estudiadas, se calculan la varianza y desviación estándar de las betas.

Para lo cual, partimos del concepto de error cuadrado medio (MSE por sus siglas en inglés), el cual se calcula de la forma siguiente:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Pero dado que dicho error cuadrado medio parte del número total de datos, existe otro indicador denominado estimador insesgado de la varianza ($\hat{\sigma}^2$ o $\hat{s}^2$), el cual ajusta el MSE de acuerdo con los grados de libertad, o lo que es lo mismo, el número de observaciones independientes. La fórmula:

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(\varepsilon)^2}{n-k} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}$$

Donde:
k = número de betas (β) calculadas.

En el caso de la regresión lineal simple, el denominador es *n –2*, porque existen dos betas, β_0 y β_1 estimadas. La fórmula quedaría:

$$\hat{s}^2 = \frac{\sum_{i=1}^{n}(\varepsilon)^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

El s^2 representa la varianza estimada de la regresión lineal simple, o cuánta variabilidad existe en la estimación completa de la regresión. Sin embargo, está en unidades cuadradas. Si recordamos, la desviación estándar se calcula como la raíz de la varianza. En econometría a la desviación estándar se le denomina error estándar (*se* o error estándar, por sus siglas en inglés). El error estándar para cada beta (β) calculada en la regresión. Las fórmulas son:

$$se(\hat{\beta}_0) = \sqrt{\left(\frac{\sum_{i=1}^{n} e_i^2}{n-2}\right)\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Significancia individual de los coeficientes *β′s*

Entre las distintas hipótesis que pueden contrastarse, las que se encuentran relacionadas con las betas (*β′s*) son de particular importancia, puesto que su aceptación o rechazo implica la aceptación o rechazo de la existencia de la relación de dependencia lineal entre las variables *x* e *y* del modelo de regresión lineal simple. Las hipótesis por comprobar se especifican de la siguiente manera:

$$H_0: \beta_0 = 0$$

$$H_0: \beta_1 = 0$$

En el caso de B_1, si rechazamos H_0 asumimos que la variable significativa. En el caso de B_0, si rechazamos H_0 asumimos que la constante es diferente de 0. Para probar las hipótesis, se utilizará el estadístico "t", que se define como $t = \frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}$ y se contrasta con el estadístico "t" de tablas $t_{n-k,\alpha}$

Al comparar con el estadístico "t" de tablas, una vez especificando los grados de libertad $n - k$, es decir, el número de observaciones (n) menos el número de coeficientes estimados (k). El estadístico "t" establece el nivel de significancia, el cual puede ser al 10%, 5% o al 1%.

Criterios

Si $t_{calculado} > t_{tablas}$ se rechaza H_0 el coeficiente de la variable $\widehat{\beta}_i$ sí es significativo al nivel α.

Si $t_{calculado} \leq t_{tablas}$ se acepta H_0 el coeficiente de la variable $\widehat{\beta}_i$ no es significativo al nivel α

Estadísticos t calculados más grandes acompañan valores p (*p values*) más pequeños.

Estadísticos t calculados más pequeños acompañan valores p (*p values*) más grandes.

Si el valor p (*p value*) asociado a la variable es menor que 0.10, se considera que la variable es significativa al 10%.

Si el valor p (*p value*) asociado a la variable es menor a 0.05, se considera que es significativa al 5%.

Si el valor p (*p value*) asociado a la variable es menor a 0.01 se considera que es significativa al 1%.

SUPUESTOS DEL MODELO DE MÍNIMOS CUADRADOS ORDINARIOS (MCO)

Una vez que hacemos un modelo econométrico y las variables son significativas, eso no es suficiente, hay que validarlo. Para validarlo se requieren ciertos supuestos, denominados de mínimos cuadrados ordinarios (MCO).

El cumplimiento de estos supuestos implica que el modelo está correctamente especificado. Dichos supuestos son la normalidad, la homocedasticidad, la especificación correcta del modelo, la ausencia de multicolinealidad y la no autocorrelación. Profundizaremos en cada uno de ellos en las secciones siguientes del presente capítulo.

Normalidad

La normalidad en un modelo econométrico implica que el error (ε) se encuentra normalmente distribuido. Esto resulta relevante, porque entonces, los estimadores de los coeficientes ($\hat{\beta}$) son estadísticamente independientes del vector de residuos y dichos residuos deberían reflejar una distribución aleatoria, es decir, sin ningún patrón aparente (Greene, 2012, pp. 42, 51; Masini y Vázquez, 2014, p. 20).

Supuesto de normalidad

En matemáticas, el supuesto de normalidad se expresa como $\varepsilon|x \sim N(0,\sigma^2 I)$ lo que significa que el error ε se distribuye de forma normal con media cero y varianza constante e independiente entre observaciones. Siendo σ^2 la varianza e I la matriz identidad de varianzas-covarianzas. Resulta clave para este supuesto que la

matriz de varianzas-covarianzas se trate de la matriz identidad (la que tiene unos en la diagonal y ceros en el resto de las posiciones de la matriz), puesto que representa la independencia y homogeneidad de la varianza de los errores.

Detección: Jarque Bera

Para identificar si el error se encuentra normalmente distribuido, se debe realizar una prueba de hipótesis en la que buscamos aceptar la H_0 de normalidad, donde *JB* representa las siglas de la prueba Jarque-Bera que se calcula de la siguiente forma:

$$JB = \frac{n}{6}\left[s^2 + \frac{(k-3)^2}{4}\right]$$

Donde:
s = representa el sesgo, *k* es la curtosis[3].

El resultado de la prueba de Jarque-Bera se compara con X_2^2, es decir con la prueba Ji cuadrada con dos grados de libertad.

Criterio: si JB < 5.99 entonces hay normalidad. Si el JB es pequeño el p value será mayor a 0.05 y se aceptará la hipótesis nula.

Regla del 6: debido a que el valor de tablas de X_2^2 con dos grados de libertad y $\alpha = 5\%$ es igual a $5.99 \approx 6$, el criterio para aceptar o rechazar H_0 es que si $JB < 5.99$, entonces se acepta H_0 y existe normalidad.

3 Las fórmulas para sesgo y curtosis son: $s = \frac{\sum(x_i-\bar{x})^3/_{n-1}}{s_x^3}$ y $k = \frac{\sum(x_i-\bar{x})^4/_{n-1}}{s_x^4}$ y en este caso las den denominador corresponden a la desviación estándar muestral, $s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$.

Corrección

Debemos partir del teorema del límite central que estipula que cualquier distribución se comporta como normal si la muestra es suficientemente grande. Es por esto que las recomendaciones para la corrección de la normalidad en los errores (ε) son:

» Incrementar la muestra.
» Eliminar datos *outliers*, es decir, las observaciones con los errores al cuadrado mayores. Una regla comúnmente utilizada es quitar los *outliers* que se encuentran a $\pm 3\sigma$ siendo σ la desviación estándar.

Homocedasticidad

El supuesto avalado por la homocedasticidad es que la varianza de los términos del error es constante en todas las características incluidas en el modelo.

Supuesto de homocedasticidad

El supuesto de homocedasticidad es en parte la profundización del supuesto para normalidad. Se había establecido que la normalidad se expresa como $\varepsilon|x \sim N(0,\sigma^2 I)$ con media igual a cero y varianza igual a $\sigma^2 I$. Entonces, suponemos que tenemos una matriz de varianza-covarianza denominada σ^2, la cual la establecemos de la siguiente forma:

Figura 3. Explicación de la homocedasticidad

$$\sigma^2 = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

y la matriz identidad es igual a

$$I = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix},$$

entonces, al multiplicarlas:

$$\sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

el resultado es:

$$\sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Fuente: elaboración propia.

Porque cualquier matriz multiplicando una matriz identidad da como resultado la matriz que se multiplicaba inicialmente. Dado que la varianza es la misma en toda la diagonal, y las covarianzas son cero fuera de la diagonal, se cumple el supuesto de homocedasticidad, o varianza uniforme. Implica una dispersión de los datos independiente de los valores de las variables del modelo.

Detección: White

Según Gujarati y Porter (2010b) y White (1980, pp. 817-818), se debe seguir el siguiente procedimiento para la detección de homocedasticidad:

1. Se lleva a cabo la regresión original (y contra las x) y se obtienen los residuos (ε) $\rightarrow Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.
2. Se elevan los residuos de la primera regresión al cuadrado y se hace una regresión contra las variables de la primera regresión, esas mismas variables al cuadrado y sus términos cruzados ($\varepsilon^2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2 + \beta_5 x_1 x_2 + \upsilon$); obtenemos R^2_{auxiliar}.
3. Se obtiene el estadístico de White → $W = n * R^2_{\text{auxiliar}}$, para contrastarlo con una X^2_{gl} con grados de libertad (gl) igual al número de regresoras sin la constante. La hipótesis nula (H_0) es Homocedasticidad; por lo que buscamos aceptar H_0.

Criterio: si W es mayor al valor X^2_{gl} crítico, entonces se rechaza la hipótesis nula y consideramos la existencia de heterocedasticidad. En cambio, si W es menor al valor X^2_{gl} crítico se acepta la hipótesis nula y consideraríamos homocedasticidad (*p-value* mayor a 0.05).

Ejemplo: se hace una regresión con 30 observaciones de Y contra x_1 y x_2 y se obtienen los residuales (ε). Después se lleva a cabo la siguiente regresión auxiliar: $\varepsilon^2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2 + \beta_5 x_1 x_2 + \upsilon$, obteniendo una R^2_{auxiliar} de 0.70. El estadístico de *White* es 30 * 0.70 = 21. Lo siguiente es contrastar el valor de 21 contra una x^2 con 5 grados de libertad, porque se realiza una regresión de 5 betas (β). En este caso $X^2_{gl=5}$. Como 21 es mayor a 11.07, se rechaza la hipótesis nula (H_0) y por lo tanto consideramos la presencia de heterocedasticidad.

Corrección

Las formas más reconocidas para ayudar en la corrección de la heterocedasticidad son las siguientes:

- Llevar a cabo una regresión de mínimos cuadrados generalizados, en lugar de mínimos cuadrados ordinarios (MCO).
- Elaborar los estimadores de White-Huber, con muestras de al menos 50 datos
- Calcular los estimadores Newey-West-Hac, los cuales son una generalización de los estimadores de White.
- Transformar los datos, siendo el método más común el cálculo de logaritmos de las variables especificadas o definir el modelo con las variables expresadas en porcentajes o razones.
- Cambiar de forma funcional, también llamado reespecificación.

Especificación correcta del modelo

Para asegurar que el modelo de MCO será capaz de predecir valores futuros de las variables independientes (x), así como de estimar la contribución individual de las variables explicativas (x) al modelo, se requiere que exista una correcta especificación del modelo econométrico.

Supuesto de especificación correcta

Las características para la selección y correcta especificación de un buen modelo son las siguientes:

- Parsimonia: entre más simple es mejor.
- Identificabilidad: se debe poder conocer qué se está estimando.
- Bondad de ajuste: el modelo especificado explica bien los datos.
- Poder de predicción: el modelo predice de forma aceptable.
- Coherencia teórica: la especificación del modelo es consistente con la teoría, sobre todo en los signos.

Asimismo, los errores más comunes de especificación son los siguientes:

- Omisión de variables: no se tienen todas las variables que ayudan a la identificabilidad del modelo. Se le conoce en inglés como *underfitting.*
- Variables irrelevantes: falla la parsimonia, se tienen variables adicionales que no se relacionan directamente con el modelo. Se le conoce en inglés como *overfitting.*
- Forma funcional errónea: en este error, falla la coherencia teórica, ya que el modelo, especialmente lo referente a los signos, no están de acuerdo con la teoría económica reconocida. Asimismo, el mejor ajuste podría ser con las variables en logaritmos en lugar de una forma funcional cuadrática o de nivel.

» Errores de medición: se pueden cometer distintos tipos de errores de medición, tanto en la variable dependiente (*y*), como en la variables o variables explicativas (*x*). Puede tratarse de registros subjetivos, incorrectos, interpolados, extrapolados o redondeados, entre otros.

Detección

Caso 1: Identificar forma funcional errónea y/o variables omitidas

De acuerdo con Gujarati y Porter (2010b), y a lo señalado por Ramsey (1969), se debe seguir el siguiente procedimiento:

1. Se lleva a cabo una regresión de la variable dependiente contra las independientes *x* y se obtiene R^2 y el pronóstico $\hat{y}$. Para fines del ejemplo suponemos solo una *x*.
2. Se hace una nueva regresión de *x* contra $x, \hat{y}^2, \hat{y}^3$; y se obtiene R^2. Si el modelo original es adecuado, las variables añadidas no serán significativas.
3. A la R^2 del primer modelo le llamaremos R_r^2 y a la del segundo modelo R_u^2.

Utilizaremos la siguiente fórmula para obtener un estadístico de contraste:

$$F = \frac{(R_u^2 - R_r^2)/m}{(1 - R_u^2)/(n - k)} \; se\ compara\ vs\ F_{m,n-k}$$

Donde:
R_r^2 = corresponde a la del modelo original.
R_u^2 = es la del modelo nuevo.
m = es la cantidad de variables regresoras nuevas (*y*).
n = cantidad de observaciones del modelo original.
k = número de estimadores (β) en el nuevo modelo.

H_0:Modelo bien especificado; entonces buscamos aceptar H_0

Criterios

Si F calculado > F de tablas entonces se rechaza la hipótesis nula y NO hay correcta especificación.

Si F calculado < F de tablas (*p-value* mayor a 0.05) entonces se acepta la hipótesis nula y habrá una correcta especificación.

Si F calculado> F de tablas (*p-value* menor a 0.05) se rechaza y, por lo tanto, el modelo está mal especificado.

Ejemplo: se lleva a cabo una regresión de *y* contra *x* con 30 observaciones, resultando en una $R_r^2 = 0.70$ Se hace una nueva regresión de contra *x* añadiendo $\hat{y}^2$, $\hat{y}^3$ de la primera regresión, y resultando una $R_u^2 = 0.73$ Se obtiene el estadístico F de la siguiente manera:

$$F = \frac{(R_u^2 - R_r^2)/m}{(1 - R_u^2)/(n - k)} = \frac{(0.73 - 0.70)/2}{(1 - 0.73)/(30 - 4)} = 1.44 \; vs \; F_{2,30-4} = 3.37$$

Al ser la F calculada < al F de tablas se acepta la hipótesis nula y, por lo tanto, podemos considerar que el modelo tiene una especificación correcta.

Caso 2: Identificar variables innecesarias

En los métodos de preprocesamiento se pueden seguir distintos pasos para la selección de variables:

» *Selección hacia adelante:* se ajusta el modelo lineal con una variable. Se van agregando variables y se llevan a cabo pruebas de hipótesis de *t* (para significancia individual) y de *F* para significancia conjunta. Se selecciona el modelo que tenga el mayor ajuste; por ejemplo, el que presente

el R^2 más alto, cuidando el principio de parsimonia (entre más simple, es mejor).

» *Selección hacia atrás:* es parecido a la selección hacia adelante. Se comienza con el modelo que integra todas las variables e identifica qué variables eliminar para mejorar los parámetros, por ejemplo, el R^2. Se da especial preferencia a eliminar las variables que no sean significativas de manera individual, y que al eliminarlas no se pierda la significancia conjunta. A este enfoque se le conoce como de Sargent-Hendry.

Corrección

Además de los test y formas de corrección vistos en la sección anterior, se pueden realizar las siguientes acciones para mejorar la especificación del modelo:

1. Cambiar la forma funcional del modelo logaritmizando la variable dependiente; alguna o todas las variables independientes; o todas las variables. Puede utilizarse el test Reset de Ramsey o el test de McKinnon-White-Davidson, el cual sirve para decidir si la forma funcional debe ser lineal o logarítmica y cuya descripción puntal se encuentra más allá de los objetivos del presente libro.
2. Para variable omitida: agregar variables relevantes o significativas, al menos de manera conjunta, como se vio en el caso 1.
3. Para errores de medición, se recomienda realizar minería de datos, reportar la fuente de donde se obtuvo la información, especificar claramente los supuestos con los que se trataron las bases de datos y/o sustituir algunas variables con variables *proxy*[4].

4 Una variable *proxy* es aquella que representa a otra variable que no puede ser observada de forma directa. En Econometría, se ha demostrado que en la estimación de variables *proxy* el signo de esta variable conserva el mismo signo que tendría la variable que no se puede observar directamente (Pratt y Krasker, 1986).

Ausencia de multicolinealidad

Para entender el porqué de la ausencia de la multicolinealidad es deseable en las regresiones econométricas, es preciso definir la multicolinealidad:

De acuerdo con Gujarati y Porter (2010b, pp. 245-248) la multicolinealidad implica una relación lineal entre las variables explicatorias , lo que conlleva a que no podemos obtener estimaciones únicas de los parámetros, lo que su vez no permite generar inferencias estadísticas, como las pruebas de hipótesis, respecto a los parámetros (β).

En economía se trabaja con identidades contables como el PIB, que se define con la fórmula: $PIB = C + I + G + X - M$. Existe multicolinealidad cuando hay identidades contables, lo cual lleva a que la significancia de los estimadores esté indeterminada.

Entonces, la presencia de multicolinealidad genera que los coeficientes o estimadores puedan parecer que están muy bien especificados, pero es debido principalmente al efecto que tiene junto a otra variable del modelo.

Por tanto, se busca activamente evitar la multicolinealidad, para asegurar que el modelo estime correctamente el efecto de cada variable en la regresión. Asimismo, la ausencia multicolinealidad permite estimar coeficientes de regresión insesgados, es decir, que se acercan a su valor verdadero.

Supuesto

La ausencia de multicolinealidad implica que ninguna (x) esté explicada ya sea parcial o totalmente por otra (x) en el mismo modelo de mínimos cuadrados ordinarios. La explicación con álgebra lineal o modelación matemática está fuera de los objetivos de material a cubrir en el presente libro.

Detección

En este caso, más que la ausencia de multicolinealidad, lo que se busca identificar es si existe multicolinealidad, para posteriormente, corregirla. Generalmente al correr un modelo de MCO que resulta

con un alto R^2 y pocos estimadores (β) significativas. Se identifica la multicolinealidad de las formas siguiente:

- Con cambios en los datos (aumento de la base de datos, diferente base de datos de prueba) los estimadores (β) cambian de signo o de nivel de significancia.
- Existe un alto R^2 y las t son pequeñas.
- La correlación entre dos variables específicas (*pairwise*) es mayor o igual a 0.5 y además dicha correlación es significativa.

Corrección

Se recomiendan las siguientes estrategias para la corrección de la multicolinealidad:

1. Eliminar información redundante, es decir, quitar las variables independientes (x) que presentan altos grados de correlación en la matriz de varianza-covarianza. Sin embargo, este método no es el más recomendado.
2. Logaritmizar las variables explicativas (x).
3. Generar indicadores como el promedio simple, el promedio ponderado o los componentes principales. Este último método es el más recomendado, ya que se conserva la mayor parte de la información, aunque se disminuye la dimensionalidad del modelo en su conjunto.

No autocorrelación

Lo que pasa con la multicolinealidad, sucede con la autocorrelación, ya que resulta más fácil entender por qué no debe presentarse en un modelo de MCO si se entiende con claridad en qué consiste. La autocorrelación se encuentra asociada con datos de series de tiempo, porque asume que el hoy depende de ayer. Entonces, lo que se busca es evitar que exista la autocorrelación en los modelos de mínimos cuadrados ordinarios.

Las causas más comunes de autocorrelación son:

» Inercia de series temporales.
» Errores de especificación.
» Efectos Cobweb, donde el pasado tiene influencia en el presente y futuro.
» Manipulación de datos, puede ser por errores o actualizaciones.

Supuesto

En el supuesto de no autocorrelación, el error (ε) en el presente del modelo de MCO no se explica con el error en el pasado ni tiene relación con los errores del futuro. Simbólicamente, significa que:

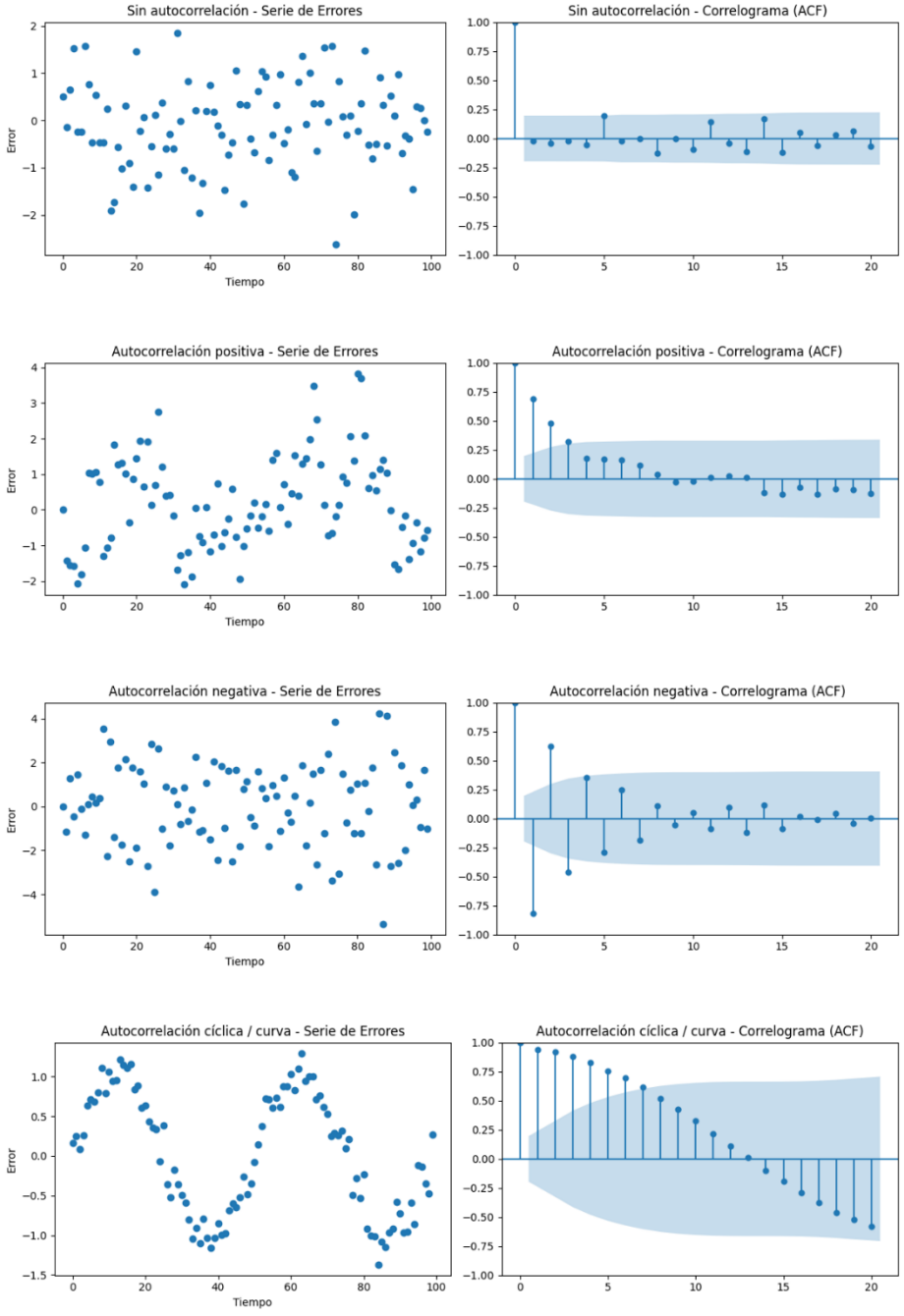
$$E(\varepsilon_i \varepsilon_j) = 0 \text{ con } i \neq j$$

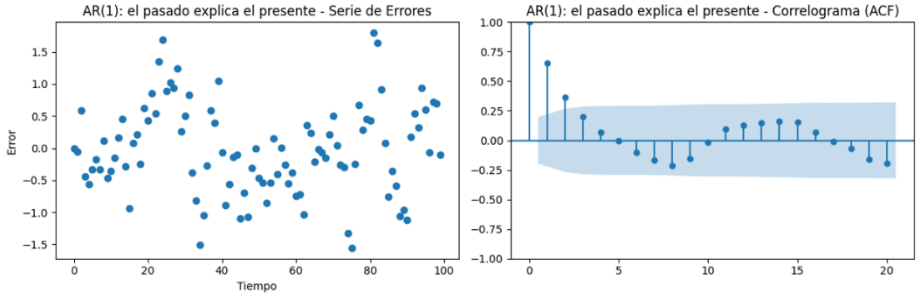
O lo que es lo mismo, el error relativo a cualquier término relacionado con una observación (del pasado, presente o futuro) no tiene relación ni está influenciada por el término del error de cualquier otra observación. En este contexto, i representa una observación dada y j otra observación distinta.

Detección

Existen dos formas para llevar a cabo la detección de la autocorrelación. El primero, es un método gráfico, que involucra graficar el error (ε) en el tiempo. Lo deseable es que los errores no presenten ningún patrón. A continuación, se muestran diversos patrones que surgen al graficar el error (ε):

Gráfica 3. Correlogramas de posibles patrones del error





Fuente: elaboración propia en Python.

El segundo método implica el cálculo del estadístico de Durbin-Watson:

$$DW = \frac{\sum_{t=2}^{t=n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{t=n}\varepsilon_t^2}$$

Implica un proceso autorregresivo de orden 1 AR(1) en el error (ε_t); en el numerador las $t \geq 1s$ porque una observación de t se pierde al realizar la diferencia respecto al período anterior ($t - 1$)

Para entenderlo, pongamos un ejemplo cercano muy sencillo. El proceso de autocorrelación en una empresa que vende pinturas y paga a sus vendedores respecto a lo que cobraron el mes anterior siempre presentará una diferencia de efectivo respecto al período anterior, y los financieros deberán prever esta necesidad desfasada de flujo de dinero. Esa relación del período actual (t) con respecto al período anterior ($t - 1$) es lo que se expresa como proceso autorregresivo de orden 1.

H_0: No autocorrelación; entonces buscamos aceptar H_0 Los pasos para realizarlo son los siguientes:

1. Se realiza la regresión de MCO y se obtienen los errores (ε_t) y ($\varepsilon_t - 1$).
2. Se calcula el estadístico de Durbin-Watson (DW), que puede tomar valores entre 0 y 4. Un valor de $DW = 2$ significa que no existe autocorrelación, $DW = 0$ denota que existe autocorrelación positiva y una $DW = 4$ es indicador de autocorrelación negativa. Para establecer valores intermedios, se tendrán que seguir los criterios de decisión para la aceptación o rechazo de la H_0

3. Se busca en la tabla del estadístico Durbin-Watson dos valores, denominados d_L y d_U. Para encontrarlos, se necesita el número de observaciones (n) y la cantidad de estimadores betas menos 1 (k – 1), que en este caso se denomina (k'). Además, se debe conocer el nivel de significancia (α) ya sea 0.05 o 0.01; d_L es el límite inferior en la distribución de Durbin-Watson (DW) y d_U es el límite superior de dicha distribución.
4. Se siguen los criterios de decisión para la aceptación o rechazo de la H_0, las cuales se explicarán a continuación.

Criterios de decisión para la aceptación o el rechazo de la H_0: por una cuestión visual, sirve dibujar en una recta numérica los valores que puede tomar el estadístico de Durbin-Watson, es decir, entre 0 y 4; siendo un valor de 0 el correspondiente a autocorrelación positiva, 2 que no existe autocorrelación y 4 indica autocorrelación negativa.

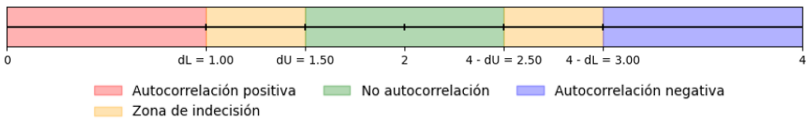
Supongamos que $d_L = 1.00$ y $d_U = 1.5$, entonces deberemos calcular 2 estimadores adicionales, $4 - d_U$ y $4 - d_L$. Con la información proporcionada, $4 - d_U = 2.5$ y $4 - d_L = 3$

Para graficar en la recta numérica se utiliza lo siguiente:

$DW = 0 \rightarrow$ *Autocorrelación positiva*
$d_L = 1 \rightarrow$ *Límite inferior del estadístico DW*
$d_U = 1.5 \rightarrow$ *Límite superior del estadístico DW*
$DW = 2 \rightarrow$ *No autocorrelación*
$4 - d_U = 4 - 1.5 = 2.5$
$4 - d_L = 4 - 1 = 3$
$DW = 4 \rightarrow$ *Autocorrelación negativa*

La recta resultante quedaría:

Figura 4. Zonas de decisión de la prueba Durbin-Watson



Fuente: elaboración propia.

Los criterios para evaluar las distintas zonas quedan como sigue:

Tabla 4. Criterios del estadístico Durbin-Watson

Criterio	¿Acepto o rechazo H_0?	Veredicto
$0 \leq DW \leq d_L$	Rechazo H_0	Autocorrelación positiva
$d_L \leq DW \leq d_U$	Zona de indecisión	Zona de indecisión
$d_U \leq DW \leq 4 - d_U$	Acepto H_0	No autocorrelación
$4 - d_U \leq DW \leq 4 - d_L$	Zona de indecisión	Zona de indecisión
$4 - d_L \leq DW \leq 4$	Rechazo H_0	Autocorrelación negativa

Fuente: elaboración propia.

Corrección

Las medidas más comunes para corregir la autocorrelación son las siguientes:

1. Estimar la regresión a través de mínimos cuadrados generalizados en diferencias, donde se transforma el modelo rezagándolo y multiplicándolo por σ, que representa la influencia del pasado $(t - 1)$ sobre el presente (t). Se realiza a través de las transformaciones de Prais-Winsten.
2. Se corrige la autocorrelación (y también la heterocedasticidad) con los estimadores de Newey-West, también denominados Newey-West HAC (Heterocedasticity Autocorrelation Correction).

Consecuencias de la violación de supuestos

Si se violan los supuestos de MCO la estimación es poco precisa; puede que una variable sea significativa cuando en realidad no lo es, y viceversa; las t y las F de las pruebas de significancia individual y

conjunta puede que no sirvan. En esta sección se especificarán las principales consecuencias de la heterocedasticidad, multicolinealidad y autocorrelación.

Consecuencias de la heterocedasticidad

Cuando se establece que existe heterocedasticidad, se asume que la varianza de los términos del error no es constante en todas las características, trayendo las siguientes consecuencias:

1. Los estimadores de MCO no son eficientes, es decir, su varianza (σ^2) no es la menor posible.
2. Existe un sesgo en las fórmulas para estimar los parámetros (β).
3. Los intervalos de confianza de *t* y *F* no son confiables en las pruebas de hipótesis, lo cual puede llevar a conclusiones equivocadas.

Consecuencias de la multicolinealidad

Las principales consecuencias de la multicolinealidad son:

1. Grandes valores en la varianza (σ^2) y la desviación estándar (σ) de los estimadores de MCO.
2. Grandes intervalos de confianza para los estimadores (β).
3. Valores de *t* muy pequeños.
4. Alto valor de R^2, pero pocas significativas.
5. Los estimadores (β) y su desviación estándar (σ) se vuelven muy sensibles a pequeños cambios en los datos, son inestables.
6. Signos erróneos en los coeficientes de la regresión. Y recordemos que es primordial que el signo de la teoría económica y el de las regresiones econométricas coincidan.
7. Dificultad para estimar la contribución individual de las variables explicativas a la suma de errores (MSE) o al R^2.

Consecuencias de la autocorrelación

Las consecuencias que se presentan en las regresiones de MCO afectadas por autocorrelación son las siguientes:

1. Al igual que en la heterocedasticidad, los estimadores de MCO no son eficientes, es decir, su varianza (σ^2) no es la menor posible.
2. Pueden generar que un coeficiente (β) sea estadísticamente significativo, sin serlo.
3. Por tanto, los test de *t* y *F* no son confiables.
4. El valor de R^2, que depende de la varianza estimada ($\hat{\sigma}^2$) no es confiable.
5. La varianza (σ^2) y la desviación estándar (σ) pueden ser ineficientes para realizar pronósticos de valores futuros.

Criterios

Si el valor p (*p value*) asociado a la prueba de normalidad es mayor a 0.05, podemos asumir normalidad en los errores.

Si el valor p (*p value*) asociado a la prueba de homocedasticidad es mayor a 0.05, podemos asumir varianza constante de los errores (homocedasticidad).

Si el valor p (*p value*) asociado a la prueba de correcta especificación es mayor a 0.05, podemos asumir especificación adecuada (variables suficientes y forma funcional adecuada).

REGRESIÓN LINEAL MÚLTIPLE

Expansión a múltiples regresores

Se parte del supuesto de que existe más de una variable independiente (x) y que además pueden tener un grado mayor a 1, es decir, las x's pueden estar elevadas al cuadrado, o al cubo; sin embargo, cumplen los supuestos clásicos del modelo MCO porque los coeficientes (β) mantienen la linealidad.

Sin embargo, para establecer una forma más general, eficiente y fácilmente derivable se utiliza un modelo que puede integrar otras formas funcionales. Se tiene:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta k\, x_k + \varepsilon$$

Donde:
y = variable dependiente.
x_1 = variables independientes con $j = 1, 2, \ldots, k$.
β_0 = intercepto.
β_j = coeficientes de las variables independientes x_j.
ε = término de error aleatorio.

Supuestos clásicos del modelo MCO

Tal y como se vieron en el capítulo 3, se llevará a cabo una recapitulación de los supuestos del modelo de mínimos cuadrados ordinarios (MCO):

1. Normalidad: implica que el error (ε) se encuentra normalmente distribuido. Se expresa como $\varepsilon|x \sim N(0,\sigma^2 I)$

2. Homocedasticidad: se relaciona con el supuesto de normalidad, en tanto que la varianza de los términos del error es constante en todas las características incluidas en el modelo. En el supuesto de normalidad de $\varepsilon|x \sim N(0,\sigma^2 I)$ la homocedasticidad es el término $\sigma^2 I$.
3. Correcta especificación: debe cumplir con la parsimonia, identificabilidad, bondad de ajuste, poder de predicción y coherencia teórica. Por otra parte se deberá evitar la omisión de variables, el incluir variables irrelevantes, especificar una forma funcional errónea o incluir errores de medición en las variables.
4. Ausencia de multicolinealidad: se busca para asegurar que el modelo estima correctamente el efecto de cada variable en la regresión, así como que permite tener coeficientes de regresión insesgados, que se acercan a su verdadero valor.
5. No autocorrelación: el error del presente no se explica con los errores del pasado ni están relacionados con los del futuro. Se expresaría: $E(\varepsilon_i \varepsilon_j) = 0$ con $i \neq j$

Evaluación de significancia

Pruebas t (significancia individual)

Al igual que en la regresión lineal simple, se utiliza la prueba para verificar la significancia. La hipótesis nula quedaría:

$$H_0: \beta_j = 0$$

Si rechazamos H_0 asumimos que la variable es significativa. Para probar las hipótesis, se utilizará el estadístico "t", que se define como $t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ y se contrasta con el estadístico "t" de tablas $t_{n-k,\alpha}$

Al comparar con el estadístico "t" de tablas, una vez especificando los grados de libertad $n - k$, es decir, el número de observaciones (n) menos la cantidad de coeficientes estimados (k). El α establece el nivel de significancia, los cuales pueden ser al 10%, 5% o al 1%.

Criterios

Si $t_{calculado} > t_{tablas}$ se rechaza $H_0 \rightarrow$ el coeficiente de la variable $\hat{\beta}_j$ sí es significativo al nivel α.

Si $t_{calculado} < t_{tablas}$ se acepta $H_0 \rightarrow$ el coeficiente de la variable $\hat{\beta}_j$ no es significativo al nivel α.

Prueba F (significancia conjunta)

H_0: $\beta_k = 0$ donde β_k representa todos los coeficientes estimados, sin incluir el intercepto. La hipótesis nula quedaría:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_j = 0$$

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \; se\ compara\ vs\ F_{k-1,n-k}$$

Donde:
$R^2 = R^2$ de la regresión no restringida.
$k =$ cantidad de coeficientes estimados (β´s).
$n =$ cantidad de observaciones.

Criterios

Si F calculado < F de tablas entonces se acepta la hipótesis nula y no existe significancia conjunta de los regresores.

Si F calculado > F de tablas (*p-value* menor a 0.05) entonces se rechaza la hipótesis nula y existe significancia conjunta.

FORMAS FUNCIONALES DE LOS MODELOS

Una vez establecidos los supuestos de los modelos lineales, tanto simples como múltiples, es importante considerar que a pesar de que un modelo de regresión sea lineal en los estimadores, no significa necesariamente que se comporte de manera lineal en las variables. Para la especificación de los modelos, se utilizan diferentes transformaciones de las variables dependientes e independientes, con lo que se pueden construir una amplia variedad de modelos, todos estimables mediante mínimos cuadrados ordinarios (MCO). A continuación describiremos algunos de éstos (Wooldridge, 2015):

Modelo lineal (nivel-nivel)

Este modelo presenta la forma funcional típica de MCO, $y = \beta_0 + \beta_1 x$ siendo β_0 el intercepto y β_1 la pendiente y se cuenta con una variable independiente y también una dependiente . En la especificación del modelo lineal o también denominado nivel-nivel no se refleja de modo explícito el término del error, pero una vez que se realice la selección de la forma funcional y se lleve a cabo la regresión a través de MCO, se tendrá que especificar el error y verificar que se cumplan todos los supuestos de MCO.

Modelo Lin-log (nivel-log)

Si lo que se busca es medir el cambio absoluto de la variable dependiente con respecto a un cambio relativo en la variable independiente (x). Entonces, se estima un modelo con la forma:

$$y = \beta_0 + \beta_1 log(x)$$

Aquí, el coeficiente de las betas (β) identifica el cambio de la con respecto a las de la siguiente manera:

$$\beta_1 = \frac{cambio\ absoluto\ en\ y}{cambio\ en\ \log(x)} = \frac{cambio\ absoluto\ en\ y}{cambio\ relativo\ en\ x} = \frac{\partial y}{\partial x/_x} = \frac{\Delta y}{(\Delta x/x)}$$

Lo que se puede interpretar es que el cambio absoluto en *y* es igual a tantas veces se multiplique la pendiente por el cambio relativo de la variable independiente (*x*). Recordando, esto es debido a que el cambio del logaritmo de un número es un cambio relativo o cambio porcentual después de multiplicarlo por 100.

Entonces, siempre que se estime este tipo de modelos, no se debe olvidar multiplicar el valor del coeficiente de la pendiente estimada por 0.01 o dividirlo entre 100. Si no se sigue este procedimiento, se podría estar llegando a conclusiones erróneas a partir de los resultados. En este modelo, la pendiente estimada se calcula como $\beta_1 \left(\frac{1}{x}\right)$.

Modelo log-lineal (log-nivel)

También llamados modelos de crecimiento o modelo semilog, ya que solo la variable dependiente aparece en forma logarítmica: $\log(y) = \beta_0 + x$ y la variable independiente se encuentra de forma lineal. En este caso, el coeficiente de las betas (*β*) miden el cambio relativo en la variable dependiente ante un cambio absoluto en el valor de la variable independiente (*x*).

$$\beta_1 = \frac{cambio\ relativo\ en\ y}{cambio\ absoluto\ en\ x} = \frac{\partial y/_y}{\partial x}$$

Para calcular el porcentaje de cambio se multiplica β_1 * 100%. A esta tasa de cambio se le conoce también como **semielasticidad** de la variable dependiente con respecto la variable explicativa o independiente.

Modelo doble logaritmo (log-log)

En este caso, tanto la variable dependiente como las dependientes se encuentran en forma logarítmica: $\log(y) = \beta_0 + \beta_1 \log(x)$ También

es llamado modelo de **elasticidad constante**, porque el coeficiente de la pendiente (β_1) puede ser interpretado como una elasticidad, esto es, como la razón del cambio porcentual de una variable dividida por el porcentaje de otra variable. Y como las elasticidades son constantes en el rango de observaciones, se denomina modelo de elasticidad constante.

Otra propiedad interesante es que la suma de los coeficientes de la regresión, es decir, las betas (β) proporcionan información sobre los rendimientos a escala, es decir, cuánto varía la y cuando cambian las variables independientes (x).

Específicamente:

- Si la suma de las betas $\Sigma \beta_i < 1$ entonces existen rendimientos decrecientes a escala, es decir, si se duplican las entradas, la salida será menos del doble.
- Si $\Sigma \beta_i = 1$, entonces existen rendimientos constantes a escala y si se duplican las entradas, se duplicarán las salidas, si se quintuplican las entradas, se quintuplicarán las salidas, y así sucesivamente.
- En cambio, si $\Sigma \beta_i > 1$, lo que sucedería es que existen rendimientos crecientes a escala, y si las entradas se duplican, las salidas serán mayor al doble que inicialmente.

Criterios para seleccionar una forma funcional

Existen varias medidas de bondad de ajuste, siendo las principales las siguientes:

1. R^2: este indicador mide la proporción de la variación explicada por los regresores. Puede tener valores entre 0 y 1. Entre más cercano esté al cero, es peor el ajuste, y entre más cercano al 1, es mejor la bondad de ajuste. Un problema es que entre más variables explicativas se incluyan en el modelo, el aumenta, por lo cual en muchas ocasiones se puede ajustar.

2. $R^2_{ajustado} = \left[1 - \frac{(n-1)}{(n-k)}\right](1 - R^2)$ el cual se usa para comparar dos o más modelos de regresión con la misma variable dependiente, pero diferente número de regresores. Dado que $R^2_{ajustado}$ es generalmente menor que R^2, al parecer penaliza el que se agreguen regresores al modelo.
3. *Criterio de información de Akaike* (*AIC*): mide el equilibrio entre la bondad de ajuste del modelo y la complejidad del modelo. Se calcula como: $AIC = 2k - 2\ln(L)$ donde k es el número de estimadores en el modelo y L es la verosimilitud del modelo. Entre más bajo es mejor. En su forma logarítmica, se define como $\ln AIC = \frac{2k}{n} + \ln\left(\frac{RSS}{n}\right)$ siendo *RRS* la suma total de los errores al cuadrado se calcula:

   $RRS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ Al comparar varios modelos, generalmente se elige el modelo con el AIC más bajo.
4. *Criterio de información de Schwarz (BIC o SIC):* también llamado Criterio de Información Bayesiana, mide lo mismo que el AIC pero penaliza más la complejidad del modelo. Se calcula $BIC = k\,ln(n) - 2\ln(L)$ y lo único nuevo que incluye es la n, que representa el tamaño de la muestra. Es un criterio alternativo al AIC, y en su forma logarítmica se expresa como sigue: $\ln BIC = \frac{k}{n}\ln n + \ln\left(\frac{RSS}{n}\right)$. En este caso, el factor de penalización es más fuerte respecto al número de estimadores k. Entre más bajo, mejor y muestra un modelo más simple (parsimonioso) con buen ajuste.

Comparación e interpretación de coeficientes

Tabla 5. Comparación e interpretación decoeficientes con diferentes formas funcionales

Modelo	**V.Dep.**	**V.Ind**	**Forma**	**Pendiente** $\frac{\partial y}{\partial x}$	**Interpretación de β_1**	
					Matemática	*Explicada*
Nivel-nivel (lineal)	y	x	$y = \beta_0 + \beta_1 x$	β_1	$\Delta y = \beta_1 \Delta x$	El cambio en "y" es igual a β_1 por el cambio en "x". Por cada unidad que cambia "x", "y" cambia β_1 La beta en este caso es una **pendiente**
Nivel-log (lin-log)	y	$\log(x)$	$y = \beta_0 + \beta_1 \log(x)$	$\beta_1\left(\frac{1}{x}\right)$	$\Delta y = (\beta_1/100)\%\Delta x$	Por cada cambio de 1% en "x", "y" cambia $\beta_1/100$ unidades.
Log-nivel (log-lin o semilog)	$\log(y)$	X	$\log(y) = \beta_0 + \beta_1 x$	$\beta_1 y$	$\%\Delta y = (100\beta_1)\Delta x$	Tasa instantánea de crecimiento. Por cada unidad de cambio en "x", "y" cambia $\beta_1 * 100\%$
Log-log (logarítmico)	$\log(y)$	$\log(x)$	$\log(y) = \beta_0 + \beta_1 \log(x)$	$\beta_1\left(\frac{y}{x}\right)$	$\%\Delta y = \beta_1 \%\Delta x$	Tasas de crecimiento. Un cambio de 1% en "x" genera un cambio de β_1% en "y" La beta en este caso es una **elasticidad**

Fuente: elaboración propia basado en Wooldridge (2015).

Tomando en cuenta que $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$

MODELOS CON DATOS DE PANEL

¿Qué son los datos de panel?

En la econometría de panel se combinan datos de series de tiempo (años, meses, semanas, días) y de corte transversal (países, ciudades, empresas, acciones, individuos). Es decir, existen dimensiones de tiempo y espacio, involucran tanto unidades de medición como unidades de tiempo. Las principales ventajas de la utilización de datos de panel son las que a continuación se establecen:

1. Permiten considerar heterogeneidad en las unidades de medición.
2. Combinan datos para aprovechar la información disponible, lo que permite que existan mayores grados de libertad (*gl*) y mejora la eficiencia.
3. Permiten estudiar dinámicas de cambio.
4. Permiten analizar conductas complejas.

El modelo general de datos de panel es el siguiente:

$$Y_{it} = \alpha_{it} + X_{it}\beta + U_{it}$$

En dicho modelo no aparece de forma explícita el término de error (ε_{it}), pero esto es debido a que se encuentra contenido en el término U_{it}. Según Mayorga y Muñoz (2000), los datos de panel se clasifican según la descomposición del error, de la siguiente manera: el error (U_{it}) es la suma de los efectos no observables que difieren entre las unidades de estudio pero no en el tiempo (μ_i) + los efectos no observables que varían entre el tiempo pero no entre

las unidades de estudio (δ_i) + el error puramente aleatorio (ε_{it}). La descomposición del error se observa en la siguiente fórmula:

$$U_{it} = \mu_i + \delta_t + \varepsilon_{it}$$

Tipos de datos panel

Lo primero que se debe observar es cómo se relacionan las observaciones, especialmente respecto a si la cantidad de datos suministrados para cada período de tiempo es el mismo o si existe alguna variación. La clasificación es como sigue:

- Paneles balanceados: la cantidad de observaciones de tiempo (t) es la misma por cada unidad de medición (i).
- Paneles desbalanceados: si la cantidad de observaciones de tiempo (t) no es la misma por cada unidad de medición (i).
  - Panel corto: tipo de panel desbalanceado en el número de unidades de medición (i) es mayor que el número de períodos de tiempo (t). Es decir, $i > t$. Un ejemplo sería observar las calificaciones de 100 alumnos durante 3 años.
  - Panel largo: tipo de panel desbalanceado en el que el número de unidades (i) es menor a la cantidad de observaciones temporales (t). Es decir, $i < t$. Un ejemplo sería observar las calificaciones de 10 alumnos durante 30 meses.

Por otra parte, existe también la posibilidad de tener paneles estáticos o dinámicos. Los estáticos se subdividen en fijos y aleatorios; los dinámicos en Arellano-Bond (GMM en diferencias) y Blundell-Bond (GMM Sytem). Veremos a continuación los modelos de panel estáticos.

Modelos pooled (datos agrupados)

Se trata de la regresión de mínimos cuadrados ordinarios (MCO), y es contra la que se verifica si conviene o no utilizar modelación de panel de datos o si es suficiente una regresión simple (o múltiple). Se utiliza cuando $\mu_i = 0$ (Mayorga y Muñoz, 2000). No se mide la

"singularidad" o heterogeneidad de las unidades de medición. De existir esta individualidad se absorbería en el término de error (ε_{it}), pudiendo estar correlacionado con las variables explicativas, generando indicadores sesgados e inconsistentes (Gujarati y Porter, 2010b). Su forma funcional sería la siguiente:

$$Y_{it} = \beta_1 + \beta_2 x_{it} + \varepsilon_{it}$$

En este modelo, se asume que todas las unidades de medición y los momentos en el tiempo son iguales, es decir, que los **datos** están **agrupados**.

Efectos fijos unidad de medición (LSDV)

Se utiliza cuando δ_t es un efecto fijo y distinto para cada unidad de medición, incorporándose esta heterogeneidad a la constante del modelo (Mayorga y Muñoz, 2000). Sirven para tomar en cuenta efectos no observables y lograr estimadores consistentes y eficientes; cada unidad, a través de una *dummy*, tiene su propio valor de intercepto (Gujarati y Porter, 2010b).

Se generan variables *dummy*[5] para cada unidad de medición, exceptuando la primera o la última, para no caer en la llamada **trampa de las dummy**, o multicolinealidad perfecta (Greene, 2012, p. 118). Por ejemplo, si tenemos 10 unidades de medición, las variables *dummy* podrían ser $m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9, m_{10}$. La *dummy* D_1 no podría generarse debido a que se pueden tener máximo $n - 1$ variables *Dummy*.

En este modelo se asumen **pendientes constantes**. El intercepto varía con las unidades de medición. Si una *dummy* no es significativa es debido a que es igual al grupo de referencia. Si las α's son significativas, entonces existen efectos asociados a la unidad de medición.

5 En econometría se especifican las *dummy* en singular a pesar de que se hable de variables en plural.

Efectos fijos unidad de tiempo (LSDV)

Se utiliza cuando es un efecto fijo para cada unidad de tiempo. Se generan variables *dummy* para todos los años excepto para el último. Por ejemplo, si tenemos 4 años las *dummy* podrían ser t_1, t_2 *y* t_3. Lo que nos interesa es que sean el número de atributos menos uno (n – 1). En este caso, las **pendientes constantes** se refieren a que el intercepto varía con el tiempo.

Efectos aleatorios (REM)

Se utilizan cuando el efecto inobservable no se correlaciona con ninguna variable explicativa (Wooldridge, 2015). En este caso, este efecto inobservable se suma al error. Esto genera que el error no sea homocedástico, por lo que se deben utilizar Mínimos Cuadrados Generalizados (MCG) para corregir esta situación (Mayorga y Muñoz, 2000).

Asume que los efectos de las variables diferentes quedan capturados en el error, el cual es un error compuesto. Su regresión resulta compleja por esta razón, y se establece de la siguiente manera:

$$Y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_{3it} + w_{it}$$

Siendo $w_{it} = \mu_i + \varepsilon_{it}$ el error compuesto de la regresión. En este se integra el error de la unidad de medición (μ_i) + el error del panel (ε_{it}). Por tanto, el modelo de efectos aleatorios quedaría:

$$Y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_{3it} + \mu_i + \varepsilon_{it}$$

Pruebas de selección del modelo

Tabla 6. Resumen de criterios para selección de modelos de panel

Elección de panel	Prueba	Hipótesis nula	Decisión
¿Pooled (datos agrupados) o efectos aleatorios?	Prueba Breusch-Pagan para efectos aleatorios	*H*0: varianza del efecto no observable $\mu i = H_0$	Si se rechaza H_0 significa que es mejor utilizar efectos aleatorios, ya que el efecto no observable sí influye en el modelo
¿Pooled (datos agrupados) o efectos fijos?	Prueba *F* de significancia de efectos fijos	H_0 : todas las variables *dummy* de las unidades de medición son igual a 0	Si se rechaza H_0 es mejor efectos fijos. Si se acepta *H*0 es mejor usar *pooled* (datos agrupados).
¿Efectos fijos o aleatorios?	Prueba de Hausman. Evalúa si existe correlación del componente aleatorio μi con las variables explicativas *X*	H_0 : los estimadores de efectos aleatorios y de efectos fijos no difieren sustancialmente (efectos aleatorios es mejor, las *dummy* no influyen)	Si se rechaza H_0 es mejor efectos fijos; y si se acepta *H*0 se escoge efectos aleatorios porque significa que no es necesario estimar las variables *dummy*
¿Incluir efectos fijos temporales?	Prueba *F* restringida	H_0: todas las variables *dummy* de las unidades de tiempo son iguales a cero	Si se rechaza H_0 implica que las variables *dummy* asociadas al tiempo deben estar en el modelo

Fuente: elaboración propia con base en Aparicio y Márquez (2005).

Prueba Breusch–Pagan para efectos aleatorios para responder a ¿pooled (MCO) o efectos aleatorios (REM)?

H_0: La regresión *pooled* (MCO) es mejor. Se calcula la prueba de Breusch-Pagan (*BP*) de la siguiente forma:

$$BP = \frac{n \cdot T}{2(T-1)}\left[\frac{\sum_{i=1}^{n}[\sum_{t=1}^{n}\varepsilon_{it}]^2 - 1}{\sum_{i=1}^{n}\sum_{t=1}^{n}\varepsilon_{it}{}^2}\right]$$

Donde:
n = cantidad de unidades de medición de corte transversal (países, ciudades, empresas, acciones, individuos).
T = número total de periodos estudiados (3 años, 30 meses).
t = índice de tiempo, se utiliza para sumar (va tomando el año 1, año 2, año 3 sucesivamente).
ε_{it} = error de la regresión.

El resultado de la prueba de Breusch-Pagan se compara con con 1 grado de libertad y se compara con:

α = 10% es igual a 2.70

α = 5% es igual a 3.84

α = 1% es igual a 6.63

Criterio

Si $BP > X_1^2$ se rechaza H_1 y entonces efectos aleatorios (REM) es mejor.

Prueba F para responder a ¿pooled (MCO) o efectos fijos (LSDV)?

H_0: La regresión *pooled* (MCO) es mejor. Se tiene que calcular la prueba *F*:

$$F = \frac{R^2_{LSDV} - R^2_{POOLED}/(n-1)}{(1 - R^2_{LSDV})/(n \cdot T - K)}$$

Donde:
R^2_{LSDV} = R^2 del modelo estimado por efectos fijos (LSDV por sus siglas en inglés).
R^2_{POOLED} = R^2 del modelo estimado por mínimos cuadrados ordinarios.
n = cantidad de unidades de medición de corte transversal (países, ciudades, empresas, acciones, individuos).
T = número total de períodos estudiados (3 años, 30 meses).
$K = \alpha + \beta$, es decir, los coeficientes de las variables *dummy* (α) + los coeficientes de las variables independientes (β). En conjunto, representan la cantidad de estimadores estimados menos la constante β_0 porque si resulta ser un modelo de efectos fijos, la constante no se incluye; pero si es un modelo *pooled* (datos agrupados), sí se toma en cuenta. Por lo que, para efectos de que no afecte, siempre se toma la cantidad de estimadores ($\alpha + \beta$) sin la constante.

El resultado de la prueba F se compara con $F_{crítico}$ de tablas, que se debe buscar como $F_{n-1,n\cdot T-K}$:

Donde:
$n - 1$ = son los grados de libertad que aparecen en el numerador de la prueba que calculamos.
$n \cdot T - K$ = son los grados de libertad del denominador.
Se debe buscar $F_{crítico}$ de tablas con un determinado α, siendo un $\alpha = 5\%$ el más común.

Criterio

Si $F > F_{n-1,n\cdot T-K}$ se rechaza H_0 y entonces efectos fijos (LSDV) es mejor.

Prueba de Hausman para responder a ¿efectos fijos (LSDV) o aleatorios (REM)?

Antes de iniciar la explicación de cómo se calcula la prueba de Hausman, es importante resaltar que lo habitual es calcularla a través de software econométrico especializado; como puede ser Stata, Eviews, Gretl, R e incluso Python.

Es por ello que el alumno no deberá preocuparse particularmente por el cálculo de dicho estadístico, ya que basta con que comprenda de manera general cómo es generado. La importancia del estudio en este apartado se centrará en la interpretación de los resultados para tomar la decisión de qué modelo es el que mejor se ajusta a los datos de panel que tiene ante sus manos. Una vez establecido este punto, se procede al establecimiento de la hipótesis nula:

H_0: La regresión de efectos aleatorios (REM) es mejor. Se tiene que calcular el criterio de Wald (*W*) para realizar la prueba de Hausman:

$$W = [\hat{\beta}_{LSDV} - \hat{\beta}_{REM}]' \psi^{-1} [\hat{\beta}_{LSDV} - \hat{\beta}_{REM}]$$

$$\psi = VARCOV(\hat{\beta}_{LSDV} - \hat{\beta}_{REM}) = 0 \rightarrow Se\ prefiere\ REM$$

$$\psi = VARCOV(\hat{\beta}_{LSDV} - \hat{\beta}_{REM}) \neq 0 \rightarrow Se\ prefiere\ LSDV$$

Para ψ, se utilizan las matrices de covarianza estimadas del estimador de pendientes del modelo de efectos fijos (LSDV) y la matriz de covarianza estimada en el modelo de efectos aleatorios (REM), excluyendo el término constante. Bajo la hipótesis nula, tiene una distribución X^2_{K-1} con $K - 1$ grados de libertad, siendo K el número de variables independientes que se están contrastando en efectos aleatorios (REM), sin incluir la constante.

Criterio

Si $W > X^2_{K-1}$ se rechaza H_o y entonces *efectos fijos* (*LSDV*) es mejor.

Prueba F restringida para responder a ¿incluir efectos fijos temporales?

H_o: los efectos fijos temporales son nulos. Se tiene que calcular la prueba restringida, la misma que calculamos en el capítulo 3, en la detección de la especificación correcta del modelo, pero cambiando los subíndices para que sean acordes a la situación actual:

$$F = \frac{(R^2_{NR} - R^2_R)/m}{(1 - R^2_{NR})/(n-k)} \; se\ compara\ vs\ F_{m,n-k}$$

Donde:
$R^2_{NR} = R^2$ de la regresión no restringida.
$R^2_R = R^2$ de la regresión restringida.
$m = R - 1$ variables *dummy* de tiempo, quitando 1 para evitar la trampa de la *dummy*.
n = cantidad de regresores sin incluir las variables *dummy* de tiempo.

Criterio

Si F calculado > F de tablas entonces se rechaza la hipótesis nula y se deben incluir los efectos fijos temporales.

Si F calculado < F de tablas (*p-value* mayor a 0.05) entonces se acepta la hipótesis nula y no se incluyen los efectos fijos temporales.

Si el valor p (*p value*) asociado a la prueba de Hausman es mayor a 0.05, asumimos que efectos aleatorios es mejor.

MODELOS DE RESPUESTA CUALITATIVA BINARIA

Naturaleza de las variables binarias y de los modelos de respuesta cualitativa

Las variables binarias son aquellas que solo pueden tomar dos valores posibles: 0 y 1, como puede ser en el caso de los resultados de un examen: aprobado (1) y reprobado (0); estados de apertura de una tienda, donde abierta se representa con 1 y cerrada con 0. Las variables *dummy*[6] se definen de la siguiente forma:

$$D_i = \begin{cases} 1 & Si\ cumple\ la\ definición\ de\ la\ categoría \\ 0 & Si\ no\ cumple\ la\ definición \end{cases}$$

Los modelos de respuesta cualitativa son aquellos en los que la variable dependiente o explicada toma el valor de 0 o de 1 (Gujarati y Porter, 2010b). Ejemplos: posibilidad de que alguien pague un crédito, de que gane un equipo, de que se pertenezca a cierto grupo, de que se dé un huracán, etcétera... Según Gujarati y Porter (2010b), también se conocen como modelos de probabilidad.

La particularidad de este tipo de modelos es que surgen de la naturaleza de los problemas estudiados, y requieren para su modelación el uso de distribuciones estadísticas como la Bernoulli, la cual da pie a la regresión logística, por ejemplo. Los detalles de cómo se llega al modelado se reservará para un esfuerzo de estudio independiente por parte del estudiante, sin embargo, se explicarán las bases

6 Recordemos que en econometría se mantiene el término *dummy* en singular a pesar de que hablemos de variables en plural.

más relevantes para la comprensión de los principales modelos de respuesta cualitativa, también llamados modelos con variable discreta. Estos modelos son calculados a través del método de Máxima Verosimilitud, el cual se explicará en una sección posterior.

Las variables *dummy* se utilizan con regularidad en los modelos de respuesta cualitativa, como pueden ser el modelo lineal de probabilidad (MLP), el modelo logit o el modelo probit. En las siguientes secciones se procederá primero a definir la máxima verosimilitud (MLE) para después proceder a especificar cada uno de los modelos con variable discreta.

Máxima verosimilitud (MLE)

Es también conocida como log-likelihood debido a que, con frecuencia, se termina estimando con su versión logarítmica para facilitar el cálculo matemático. ¿Pero qué es? Es una función de los estimadores condicionada a los datos proporcionados para la regresión. En español, significa que cuando tienes una función de probabilidad y no conoces los parámetros (β) los estimas por máxima verosimilitud.

Según Greene (2012, pp. 468-469), la función de máxima verosimilitud, que se estima de la siguiente manera:

$$L(\theta|y) = f(y_1, \dots, y_n|\theta) = \prod_{i=1}^{n} f(y_i\,|\theta)$$

Sin embargo, así es muy difícil estimarla porque se trata de una multiplicatoria ($\prod$) y la θ representa la recopilación de los datos muestrales. Por lo que, para facilitar el maximizarla, se toma el logaritmo natural de ambos lados:

$$Ln\ L = Ln \prod_{i=1}^{n} f(y_i\,|\theta)$$

Lo cual puede representarse también como:

$$\ln L(\theta|y) = \sum_{i=1}^{n} \ln f(y_i|\theta)$$

Si asumimos que los errores del modelo $f(y_i|\theta)$ siguen una distribución normal[7], entonces el log-likelihood queda de la siguiente manera:

$$\text{Ln } L = -\frac{1}{2}\sum_{i=1}^{n}\left[\ln \sigma^2 + \ln(2\pi) + \frac{(y_i - [\beta_0 + \beta_1 x_i])^2}{\sigma^2}\right]$$

Dado que ya definimos que el log-likelihood sigue una distribución normal, el parámetro θ se integra en la función, la cual se presenta factorizada. Para este caso, en que se asume que los errores del modelo siguen una distribución normal, a partir de la función de máxima verosimilitud (MLE) el estimador $\hat{\beta}_{MLE}$ es igual al estimador de MCO $\hat{\beta}_{MCO}$, que justo es lo que se está buscando, ya que a pesar de que se trata de variables binarias, los estimadores son iguales.

Modelo lineal de probabilidad (MLP)

Es el más sencillo. Simplemente se lleva a cabo una regresión de mínimos cuadrados ordinarios (MCO). Sin embargo, esto presenta varios problemas (Gujarati y Porter, 2010b):

- Los errores no son normales.
- La varianza de los errores no es homocedástica.
- No hay garantía de que la probabilidad resultante esté entre 0 y 1.
- No se considera confiable el R^2 en este tipo de modelos.
- Las probabilidades en la realidad no se comportan de manera lineal.

Las soluciones posibles a dichos problemas se enlistan en el mismo orden que los problemas:

- Se generan muestras grandes.
- La heterocedasticidad se puede corregir con mínimos cuadrados ponderados.

7 La ecuación de la función de densidad de probabilidad de la distribución normal se representa así: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

- » Se puede solucionar suponiendo que valores menores de 0 sean igual a 0, y valores mayores a 1 sean igual a 1; o re escalando.
- » En este caso, se utilizan otros indicadores para la evaluación de los modelos.
- » Se utilizan otro tipo de distribuciones como la logística y la normal acumulada, que pueden solucionar el problema de que la probabilidad no se encuentre entre 0 y 1.

El modelo de probabilidad lineal se define de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Donde:

$Y_i \begin{cases} 1 & P_i \textit{ Éxito} \\ 0 & 1 - P_i \textit{ Fracaso} \end{cases}$ es decir, corresponde a la variable dicotómica.

X_i = corresponde a las variables independientes.
ε_i = término de error.

En este caso particular, la estimación se lleva a cabo por la metodología de mínimos cuadrados ordinarios (MCO).

Modelo logit

La variable binaria , al igual que en el modelo de probabilidad lineal, se define como:

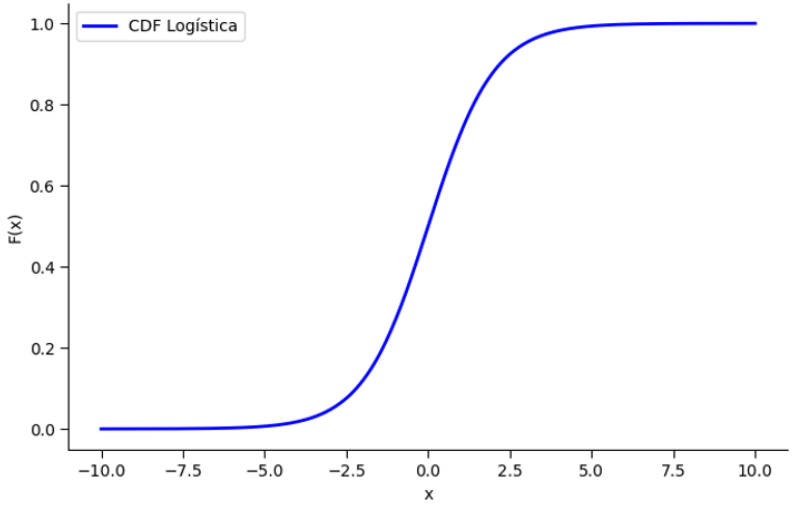
$$Y_i \begin{cases} 1 & P_i \textit{ Éxito} \\ 0 & 1 - P_i \textit{ Fracaso} \end{cases}$$

El modelo logit inicial se buscaría estimar con la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_i$$

Pero la función acumulativa de la distribución logística (logit) se representa de la siguiente manera:

Gráfica 4. Función de distribución acumulada logística



Fuente: elaboración propia.

$$P_i = E[Y_i = 1|x_i] = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

Es decir, la ecuación evalúa la probabilidad de éxito, dado que en el valor correspondiente de la variable explicativa . Para llevar a cabo la derivación del log-likelihood, debemos expresar como un indicador, es decir:

$$z = \beta_0 + \beta_1 x_i$$

Por lo que la función logística se re expresaría como:

$$P_i = \frac{1}{1 + e^{-z}}$$

Si buscamos despejar e^{-z} se multiplican ambos lados de la expresión $\frac{1}{1+e^{-z}}$ por $[1+e^{-z}]$ para eliminar el denominador, quedando de la siguiente manera:

$$[1 + e^{-z}]P_i = \frac{1}{1 + e^{-z}}[1 + e^{-z}]$$

$$[1 + e^{-z}]P_i = 1$$

$$[1 + e^{-z}] = \frac{1}{P_i}$$

$$e^{-z} = \frac{1}{P_i} - 1 = \frac{1 - P_i}{P_i}$$

$$e^{-z} = \left[\frac{1 - P_i}{P_i}\right]^1$$

Pero como normalmente se expresan los exponentes de forma positiva, e^{-z} quedaría:

$$e^{z} = \frac{P_i}{1 - P_i}$$

Esta expresión se denomina **razón de momios**. La definición más sencilla es que $\frac{P_i}{1-P_i}$ y representa el cambio de probabilidad del éxito respecto al fracaso. Por ejemplo, si $P_i = 0.9$, entonces la probabilidad de fracaso $1 - P_i = 0.1$; la razón de momios es $\frac{P_i}{1-P_i} = \frac{0.9}{0.1} = 9$, lo que querría decir que hay una probabilidad de 9 a 1 a favor de tener éxito.

Pero, aquí no acaba, porque la expresión de máxima verosimilitud para la regresión logística[8] para estimar los coeficientes (β) se podría expresar como:

$$L(\beta) = \prod_{i=1}^{n} P_i^{Y_i}(1 - P_i)^{1-Y_i}$$

Sustituyendo:

$$P_i = \frac{1}{1 + e^{-z}}$$

La expresión de máxima verosimilitud quedaría:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{1}{1 + e^{-z}}\right]^{Y_i} \left[1 - \frac{1}{1 + e^{-z}}\right]^{1-Y_i}$$

Ahorrándonos la derivación matemática, si asumimos que los errores del modelo siguen una distribución logística, entonces el log-likelihood queda de la siguiente manera:

$$Ln\ L(\beta) = \sum_{i=1}^{n} [Y_i z - \ln(e^z + 1)]$$

Con:

$$z = \beta_0 + \beta_1 x_i$$

$$Ln\ L(\beta) = \sum_{i=1}^{n} \left[Y_i(\beta_0 + \beta_1 x_i) - \ln\left(e^{\beta_0 + \beta_1 x_i} + 1\right)\right]$$

8 Se identifica de esta manera ya que la distribución logística parte de la distribución de Bernoulli condicional a x, cuya función se establece como $f(y_i|\beta) = P_i^{Y}(1\text{-}P_i)^{1\text{-}Y}$ y se parte de que la probabilidad de que sucedan dos eventos a la vez está dado por $P(A \cap B) = P(A) \cdot P(B)$.

La interpretación directa de los coeficientes no nos dice mucho, son pendientes parciales. En el caso del modelo logit se profundizó y se desarrollaron de forma tan completa las ecuaciones porque este modelo es la base de la regresión logística en *machine learning* y del aprendizaje profundo de redes neuronales (*deep learning*), las cuales tienen aplicaciones múltiples en inteligencia artificial (IA). De forma particular para este tipo de modelos, cabe resaltar que la denominada función *cross-entropy* se deriva del término:

$$Ln\ L(\beta) = \sum_{i=1}^{n} [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)]$$

Modelo probit

En este modelo lo que estimamos es el valor de Z de la distribución normal (O'Halloran, 2005a; 2005b). Altos valores de Z Es decir, cuál es el porcentaje de la normal acumulada hasta el estadístico Z.

Debido a esto, la probabilidad queda como sigue:

$$P_i = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ dz$$

Pero, dado que se estandariza con $\mu = 0$ y $\sigma^2 = 1$, con $z_i = \beta_0 + \beta_1 x_i$ quedaría:

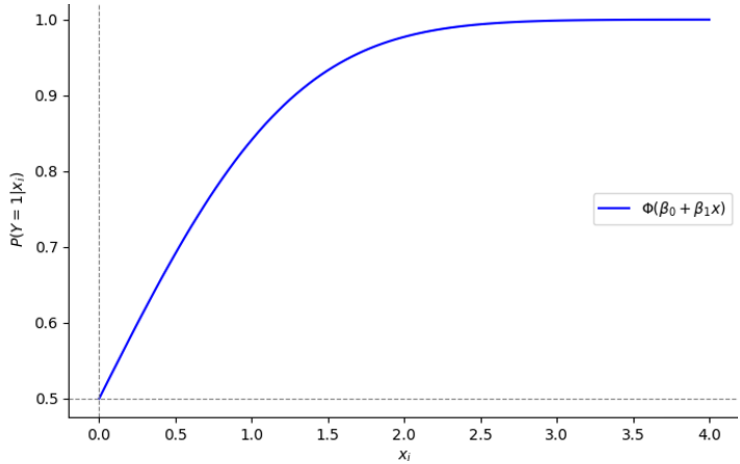
$$P_i = \int_{-\infty}^{\beta_0+\beta_1 x_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\ dz$$

La cual también se puede establecer de forma más compacta (que nos servirá al estimar MLE) de la siguiente manera:

Visualmente:

$$P_i = \Phi(z_i)\ o\ también\ P_i = \Phi(\beta_0 + \beta_1 x_i)$$

Gráfica 5. Modelo probit: probabilidad estimada de Y = 1



Fuente: elaboración propia.

El estimador de verosimilitud en su versión de log-likelihood es el siguiente (Greene, 2012, pp. 710-735; O'Halloran, 2005a, p. 61):

$$\ln L(\beta) = \sum_{i=1}^{n} \left[Y_i \ln \Phi(z_i) + (1 - Y_i) \ln\left(1 - \Phi(z_i)\right) \right]$$

Bondad de ajuste en modelos cualitativos

La R^2 tradicional no sirve para determinar la bondad de ajuste en este tipo de modelos. Existen varias medidas alternativas. Una de ellas es la del porcentaje correctamente predicho, que no es más que la proporción de veces que el modelo acierta (Wooldridge, 2015). Se considera acierto si el valor de la variable dependiente (y) es 1 y su pronóstico es ≥ 0.5; o también si la variable dependiente es 0 y el pronóstico es < 0.5. En otro caso se considera que no hay acierto.

Función de predicciones correctas (FPC)

Asume que la probabilidad del modelo depende del número de aciertos, recordando que el criterio de 0.5 es el que ayuda a definir el éxito (P_i) o fracaso ($1 - P_i$). Se calcula como sigue:

$$FPC = \frac{\#\ aciertos}{n}$$

Criterio

$$Acierto \begin{cases} Y_i = 1\ y\ P(Y_i = 1|x_i) > 0.5 \\ Y_i = 0\ y\ P(Y_i = 1|x_i) < 0.5 \end{cases} \quad Fracaso \begin{cases} Y_i = 1\ y\ P(Y_i = 1|x_i) < 0.5 \\ Y_i = 0\ y\ P(Y_i = 1|x_i) > 0.5 \end{cases}$$

Pseudo R^2 de McFadden

Según Greene (2012, p. 683), funciona de manera similar al R^2 en cuanto a que las "x" explican el fenómeno de interés; pero casi no se usa en la práctica.

$$Pseudo\ R^2 = 1 - \frac{\ln L}{\ln L_0}$$

Siendo ln$L$ el log-likelihood de la regresión completa y lnL_0 solo considerando el intercepto, ya que no se toman en cuenta las "x". Este pseudo R^2 tiene valores entre 0 y 1. Si todos los coeficientes de la pendiente son cero, es igual a cero. No es posible que sea 1, pero se puede acercar bastante.

Pruebas de hipótesis múltiple para modelos logit y probit

Se usa el llamado estadístico de razón de verosimilitud (*LR*), también llamada máxima verosimilitud.

H_0: Modelo restringido es mejor. Se calcula *LR*:

$$LR = 2(L_U - L_R)$$

Donde:
L_U = Valor de máxima verosimilitud sin restricciones, considerando todas las variables.
L_R = Valor de máxima verosimilitud restringido.

El resultado de la prueba *LR* se compara con X_1^2, es decir un X^2 con 1 grado de libertad y se compara con:

$$\alpha = 10\% \text{ es igual a } 2.70$$

$$\alpha = 5\% \text{ es igual a } 3.84$$

$$\alpha = 1\% \text{ es igual a } 6.63$$

Criterio

Si $LR > X_1^2$ se rechaza H_0 y entonces el Modelo Restringido (L_R) es mejor.

Se debe buscar X_1^2 de tablas con un α determinado α= 5%, siendo un el más común.

Factores de conversión de Amemiya

Según Amemiya, (1981), sirven para hacer comparables los resultados del modelo lineal de probabilidad, el modelo logit y el modelo probit. Se trata de llevar a cabo algunas conversiones:

Tabla 7. Factores de conversión entre logit, probit y MLP

Factor de conversión 1(β_k)	$\hat{\beta}_{LOGIT} \approx \hat{\beta}_{PROBIT} \left(\frac{\pi}{\sqrt{3}}\right)$
Factor de conversión 2(β_k)	$\hat{\beta}_{LOGIT} \approx \frac{\hat{\beta}_{PROBIT}}{5/8} \approx \frac{\hat{\beta}_{PROBIT}}{0.625}$
Factor de conversión 3(β_k)	$\hat{\beta}_{MLP} \approx \hat{\beta}_{LOGIT}\ (0.25)$
Factor de conversión 4(β_o)	$\hat{\beta}_{MLP} \approx \hat{\beta}_{LOGIT}\ (0.25) + 0.5$

Fuente: elaboración propia.

ECONOMETRÍA ESPACIAL

Introducción a la econometría espacial

El análisis espacial busca una explicación para los diversos fenómenos geográficos, a través de la conformación de estructuras y relaciones que se llevan a cabo en un territorio determinado, y evaluando el comportamiento de entidades, ya sean empresas o personas, en un espacio geográfico delimitado.

Inicialmente, se deberá llevar a cabo un análisis exploratorio de los datos espaciales (AEDE), es decir, la información geográfica que contiene información de latitud y longitud. La latitud está definida como el eje Y, cuya referencia es el meridiano de Greenwich. En cambio, la longitud, tiene que ver con el eje X, y su punto inicial es el Ecuador.

Los principales análisis exploratorios son:

- **Mapas temáticos**, resaltando una variable a la vez; en el mapa las variables se representan utilizando colores, cuya intensidad varía según el cuantil al que pertenezca una región determinada.
- **Histogramas de frecuencia** de las variables de interés, se representa el histograma junto con un mapa en que se represente por colores las distintas frecuencias.
- **Diagrama de caja y bigote** (*Boxplot*), en el que se genera el Boxplot junto con el establecimiento de los datos atípicos, lo cuales se verán representados en mapas.
- **Diagramas de dispersión** de las variables; se grafica la variable endógena del modelo y cada una de las exógenas. Da como resultado una aproximación del comporta-

miento de cada variable exógena, estableciendo qué tan explicativa es de la endógena.

La principal información necesaria para generar un análisis territorial es la siguiente:

» Datos externos: es toda la información que obtenemos principalmente de fuentes públicas. Pueden tratarse de organizaciones públicas o privadas, nacionales o internacionales. Ejemplos de este tipo de instituciones son el Banco Mundial, la ONU, el Instituto de Información Estadística y Geográfica de Jalisco (IIEG) o el INEGI.
» Datos internos: las compañías, en la operación del negocio, generan una información propia de su funcionamiento: financiamiento, transacciones, pedidos, contratos, envíos, facturas, letras de cambio, datos personales de los clientes.
» Cartografía digital: se requiere información de mapas para llevar a cabo el análisis exploratorio inicial. La cartografía puede encontrarse en distintos formatos, siendo los más comunes el Raster; datos vectoriales como archivos .shp (geometría), .shx (archivos índice), .dbf (tabla o base de datos), .prj (información de proyección); GeoPackage, entre otros.
» Datos geográficos: consiste en la representación digital de un dato espacial o rasgo geográfico que puede referirse espacial y temporalmente, dotado de atributos como son la latitud y la longitud.

Los Sistemas de Información Geográfica (GIS, por sus siglas en inglés) están preparados para geocodificar bases de datos y representarlas en mapas y son la principal forma en que se lleva a cabo el análisis inicial de los datos espaciales.

Modelo básico de regresión lineal (MBRL)

Se trata del modelo econométrico cuyas unidades de observación son intrínsecamente geográficas, es decir, pueden ser representadas en mapas y georreferenciadas a través de datos geográficos, con latitud y longitud; como pueden ser países, estados, municipios, áreas geoestadísticas básicas (AGEB), entre otros.

En el modelo básico de regresión lineal, se cumplen todas las hipótesis básicas para estimar por mínimos cuadrados ordinarios (MCO), sin embargo, se asume que presenta autocorrelación espacial. Para verificarlo, se valida que el efecto espacial se encuentre explicado por los valores de una o más variables explicativas (x). Por lo que se incluye en el modelo un número (K) de variables explicativas de tal manera que se produce una ausencia de significancia en la relación espacial entre y_i e y_j, de tal manera que $Cov\,(y_i\,, y_j) = 0$.

Los modelos más comúnmente utilizados en econometría espacial se especifican en las siguientes secciones (Anselin, 1988; Chasco, 2012).

Modelo de error espacial (SEM)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \; \forall i = 1,2, \ldots n$$

Donde:
y_i = variable dependiente para la unidad i.
x_{ij} = valor de la variable independiente en la unidad i.
β_j = coeficiente estimado para el regresor j.
ε_i = término del error para la unidad i.

Siendo el error definido de la siguiente manera:

$$\varepsilon_i = \lambda \sum_{j=1}^{n} w_{ij} \varepsilon_j + u_i$$

Donde:
w_{ij} = matriz de pesos espaciales.
λ = estimador de corrección espacial en los errores.
u_i= término de error.

Modelo espacial autorregresivo (SAR)

En este caso se asume que las observaciones tienen un lag espacial. Se define como:

$$y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Donde:
ρ = intensidad del efecto espacial (lag espacial).
w_{ij} = matriz de pesos espaciales, representa la cercanía entre las unidades espaciales.
y_j = variable dependiente en la unida , que es vecina de .
$x_{i1}, \ldots, x_{ik}$ = variables independientes de la unidad geográfica .
ε_i = error aleatorio.

Matriz de pesos espaciales (w_{ij})

En los modelos espaciales, las interacciones entre los distintos datos geográficos se representan mediante la matriz de pesos espaciales. El elemento de los subíndices *i, j* describen la proximidad entre la unidad *i* y la unidad *j* en términos de distancia. Si $w_{ij} \neq 0$, entonces *j* es vecina o contigua a la unidad *i*; pero si $w_{ij} \neq 0$, entonces *j* no es vecina de la unidad *i*.

Considerando el espacio "A", los tipos de matrices de pesos espaciales de acuerdo con la contigüidad o vecindad pueden ser:

Queen contiguity o criterio de la reina: para crear la matriz de pesos espaciales se considera que dos municipios serían vecinos según este criterio si comparten un borde, frontera o un vértice de sus respectivos límites territoriales.

1	1	1
1	A	1
1	1	1

Rook contiguity o criterio de la torre: en este caso dos municipios se consideran vecinos solo si tienen bordes territoriales (pero no vértices) en común.

0	1	0
1	A	1
0	1	0

Distance weight: a partir de las coordenadas XY de los centroides de los municipios, se obtendría una matriz de distancias entre todos los municipios. No obstante, tenemos la opción de especificar un punto de corte para determinar la distancia mínima a partir de la cuál consideramos vecinos a dos municipios.

K-Nearest Neighbors: con este criterio debemos especificar el número de vecinos que queremos que tenga cada municipio, de forma que en la matriz resultante se considerarían vecinos solo los "K" indicados más próximos. Se calcula el centroide de los polígonos y se define cuál es el K vecino más cercano con base en la distancia lineal del centroide de interés a los vecinos.

Análisis de autocorrelación espacial

En los fenómenos espaciales, la presencia de autocorrelación es esperada, puesto que esta ayuda a identificar los patrones que presentan los datos espaciales en un territorio estudiado. Con el índice I de Moran y el índice LISA se identifica de forma matemática la autocorrelación espacial, tanto a nivel global como local.

I de Moran global (Anselin)

Según Anselin, (1988), es una medida de la autocorrelación espacial de manera global, basándose en la información de las características geográficas estudiadas, evaluando si hay un patrón de autocorrelación que se presenta de manera dispersa, agrupada o aleatorio. En el último caso, se establece que no existe correlación, ya que un patrón

aleatorio desmiente la autocorrelación espacial. El índice está definido de la siguiente forma:

$$H_0 = No\ existe\ autocorrelación\ espacial$$

$$I = \frac{n}{S_0 \sum_{i=1}^{N} z_i^2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} z_i z_j$$

$$s_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$$

$$z_i = (x_i - \overline{x})$$

Donde:
$z_i = (x_i - \bar{x})$ es la diferencia del atributo geográfico para la característica *i* respecto a su media.
$z_j = (x_j - \bar{x})$ es la diferencia del atributo geográfico para la característica *j* respecto a su media.
w_{ij} = corresponde a la matriz de pesos espaciales entre el objeto espacial *i* y el objeto espacial *j*.
N = cantidad total de características.
S_0 = suma de los pesos espaciales.

El resultado de *I* se encuentra entre -1 y 1. Además, se debe tomar en cuenta el signo de z_i y el *p-value* de *I* para establecer el criterio de interpretación.

Criterio

Si el *p-value de I* resulta no significativo se puede rechazar H_0 y entonces no hay evidencia de autocorrelación, por tanto, se asume que el patrón geográfico es aleatorio.

Si el *p-value de I* resulta significativo, se rechaza H_0 y se procede a identificar el signo de z_i, presentándose dos escenarios:

- » $z_i > 0$ Existe una distribución agrupada de las zonas geográficas estudiadas.
- » $z_i < 0$ Existe una distribución dispersa de las zonas estudiadas.

Local Indicators of Spatial Association (LISA)

Según Anselin, (1995), la relación de agrupamiento ya sea concentrada o dispersa, establecida por el índice I de Moran, nos habla de que existe autocorrelación, sin embargo, no explica exactamente cómo se presentan los patrones espaciales; si con una alta concentración de unidades o baja concentración. Para lograr la especificación de dichos patrones, se usa LISA.

Según Anselin (1995), dicho índice proporciona información sobre los valores atípicos de la correlación espacial, así como ayuda a la identificación de la localización de clústeres, ayudando a la clasificación de los mismos. Se calcula de la siguiente manera:

$$I = z_i \sum w_{ij} z_j$$

Donde:
$z_i = (x_i - \bar{x})$ es la diferencia del atributo geográfico para la característica respecto a su media.
$z_j = (x_i - \bar{x})$ es la diferencia del atributo geográfico para la característica respecto a su media.
w_{ij} corresponde a la matriz de pesos espaciales entre el objeto espacial *i* y el objeto espacial *j*, estableciendo por convención que $w_{ij} = 0$.

El resultado de *I* se encuentra entre -1 y 1, al igual que en el I de Moran. Según Chasco (2006), se establece la significancia de cada agrupamiento, de acuerdo con los criterios que se mencionarán más adelante. Por lo general, se calcula primero el I de Moran y posteriormente el LISA. El I de Moran te dice si existe autocorrelación espacial y LISA qué tipo de relación existe en ese agrupamiento. El LISA suele presentarse en el software de acuerdo con colores predeterminados, los cuales se replican para la explicación que a continuación se establece. Los colores se pueden cambiar en caso de requerir colores con diferentes niveles de contraste o tonalidades.

Tabla 8. Interpretación de los conglomerados del índice LISA

Clúster	Color	Descripción
High-High		El clúster presenta un valor superior al promedio y se encuentra circundado por clústeres que también están por encima del promedio de la variable estudiada.
High-Low		El clúster presenta un valor superior al promedio, pero se encuentra rodeado de clústeres con valores por debajo del promedio respecto a la variable estudiada.
Low-High		El clúster posee un valor por debajo del promedio, pero se encuentra circundada por clústeres con valores por encima del promedio respecto a la variable estudiada.
Low-Low		El clúster posee un valor por debajo del promedio y los clústeres que lo rodean también presentan valores por debajo del promedio de la variable estudiada.
No significativo		El valor de la variable estudiada del clúster no se relaciona de forma significativa con los valores de los clústeres vecinos.

Fuente: elaboración propia con base en la descripción del software Arcgis (ArcGis Pro, 2023).

Recomendaciones de software (GeoDA, Python, Otros)

Casado et al. (2012) establecen que un Sistema de Información Geográfica (SIG O GIS) es un software con herramientas para reunir, introducir, almacenar, recuperar, transformar y cartografiar datos espaciales para un conjunto particular de objetivos propuestos.

De acuerdo a los mismos autores, entre las funciones principales de un GIS está el permitir la entrada de información, para convertir la información geográfica analógica a mapas digitales; la representación gráfica y cartográfica, al mostrar los resultados de operaciones sobre los mapas digitales introducidos previamente; la administración de información espacial, extrayendo y reorganizando información geográfica; y, finalmente, funciones analíticas, como puede ser el procesamiento e integración de datos para obtener información de tendencias o modelos espaciales.

GeoDA

En principio, se recomienda el uso del software gratuito GeoDA para el análisis básico de las bases de datos con información georreferenciada, el cual sirve como introducción a la ciencia de datos espaciales. Está diseñado para facilitar nuevos conocimientos a partir del análisis de datos mediante la exploración y modelado de patrones espaciales.

GeoDa fue desarrollado por el doctor Luc Anselin (2023) y un equipo de expertos en análisis espacial. Este software combina la funcionalidad de análisis espacial y econometría, ya que cuenta con un menú de regresión, y sirve para la realizar el análisis exploratorio de datos espaciales (AEDE).

Desde su lanzamiento inicial en febrero de 2003, el número de usuarios de GeoDa ha aumentado exponencialmente a más de 520,000. Esto incluye a los usuarios de laboratorios de universidades como Harvard, MIT y Cornell.

GeoDa se ejecuta en Windows, MacOSX y Linux (Ubuntu). Se debe ingresar al siguiente enlace: https://geodacenter.github.io/ se debe entrar en la pestaña denominada "Download" y seleccionar el sistema operativo de su preferencia.

Tabla 9. Generalidades para introducir datos en GeoDA

Elementos de archivos shp	Contenido	Archivos
.shp	Geometría	.xls
.shx	Archivo índice	.xls
.dbf	Tabla o base de datos	.txt, .doc, .xls
.prj	Información sobre los datos de proyección	.mdb

Fuente: elaboración propia.

Es necesario contar con al menos tres de los archivos que integran un shape (shp) correctamente referenciado y con proyección geográfica adecuada.

Para un análisis más específico de la dependencia o autocorrelación espacial, GeoDa ofrece otras herramientas útiles que parten de la construcción de matrices de interacciones espaciales, o expresado de una forma más general, de matrices de pesos espaciales.

Si lo que deseamos es conocer si existe relación entre el comportamiento de un municipio y el de sus vecinos, lo primero que tenemos que abordar es la definición de vecindad, es decir, cuándo consideramos que dos municipios son vecinos.

Tabla 10. Municipios vecinos

	Mun 1	Mun 2	Mun3	Mun 4	...	Mun n
Mun 1	0	1	1	0	0	0
Mun 2	1	0	0	1	0	0
Mun 3	1	0	0	0	1	0
Mun 4	0	1	0	0	0	1
...	0	0	1	0	0	1
Mun n	0	0	0	1	1	0

Fuente: elaboración propia.

Las celdas correspondientes a municipios vecinos tomarían el valor 1, mientras que el resto de la matriz sería igual a 0. Es importante señalar que en este caso la diagonal es cero porque un municipio no puede ser vecino de él mismo. Se puede observar que, en el caso particular de la matriz observada como ejemplo, cada municipio tiene exactamente dos vecinos y la matriz es simétrica.

Python

Python es un lenguaje de programación Open Code que permite integrar sistemas de forma eficaz. Es potente y posee una sintaxis clara, que a pesar de no tratarse de "lenguaje natural" como el que se utilizaría con una Inteligencia Artificial, permite entender el código de forma sencilla.

Al tratarse de un lenguaje multipropósito, es importante identificar las librerías qué se deben utilizar dependiendo el objetivo buscado. En el caso que nos atañe, describiremos las principales librerías que se utilizan en econometría y econometría espacial:

» Statsmodels: es un desarrollo de Jonathan Taylor, quién lo desarrolló y mejoró posteriormente durante el Google Summer of Code 2009. Se actualiza constantemente y se comparan los resultados con R, Stata o SAS. Se instala en Colab de Google con la siguiente instrucción:

```
pip install statsmodels
```

» Linearmodels: en esta librería se desarrollan modelos que no se encuentran en Statsmodels, como modelos de panel, modelos con variables instrumentales, estimadores de sistemas de regresiones y modelos de factores lineales. La instalación en Colab de la librería completa se realiza así:

```
pip install linearmodels
```

» PySal: esta librería es muy interesante, ya que fue desarrollada por un equipo liderado por Anselin, el mismo autor de los Análisis de Autocorrelación Espacial I de Moran y LISA, estudiados previamente en el presente capítulo. También es parte del equipo que desarrolló GeoDA, por lo cual se garantiza una coherencia teórica y metodológica con el uso de esta librería. Su instalación en Colab es como sigue:

```
pip install pysal
```

» Geopandas: extiende las capacidades de la librería pandas permitiendo la manipulación de datos geográficos en Python. Facilita operaciones como la carga, manipulación y análisis de datos espaciales con estructuras de datos similares a las de pandas. Pandas, por su parte, genera estructuras de datos rápidas, las visualizaciones son esté-

ticas y es ideal para manipular y analizar datos de entrada. La instrucción para instalar Geopandas queda:

```
pip install geopandas
```

» Folium: Facilita la creación de mapas interactivos basados en la biblioteca JavaScript Leaflet.js. Permite superponer capas, marcas y widgets sobre mapas para análisis visuales detallados.

```
pip install folium
```

» Polygon: pertenece a la librería shapely.geometry y se usa para definir y manipular figuras geométricas de dos dimensiones, permitiendo realizar operaciones como calcular el área, el perímetro y otros atributos relacionados. Se instala de una forma distinta, porque Polygon está dentro de shapely.geometry:

```
from shapely.geometry import Point, LineString, Polygon
```

» Geohexgrid: es una librería que sirve para trabajar con hexágonos en un sistema de mapeo. Muy útil para crear y manipular mallas hexagonales sobre mapas. Su instalación se realiza con el siguiente comando en Colab:

```
pip install geohexgrid
```

Otros softwares para el análisis espacial

En el mercado existen opciones para llevar a cabo análisis espacial en general (sobre mapas) y de econometría espacial. La plataforma tecnológica líder en sistemas de información geográfica (GIS) es ArcGIS, la cual se encarga de integrar y conectar datos geográficos. Es un software de paga capaz de crear, gestionar, analizar, mapear y compartir todo tipo de datos.

Existen versiones que emulan la funcionalidad de ArcGIS (sin igualarla), pero que pueden ser descargados de forma gratuita,

como QGIS (https://www.qgis.org/) el cual permite visualización espacial, principalmente.

El INEGI desarrolló su propio GIS, el Mapa Digital de México (INEGI, 2025), el cual es un software con herramientas para la construcción, consulta, interpretación y análisis de la información geográfica y estadística con georreferenciación.

La gran ventaja de este mapa digital es que toda la información al utilizar el software es congruente, generalmente no se requiere llevar a cabo proyección de datos geográficos, especialmente de archivos tipo shape. La mayor parte de la información económica, de población y vivienda y de las encuestas tienen integrados archivos que se pueden cargar directamente a esta herramienta.

La principal "desventaja" es que no se trata de un software intuitivo, sino que requiere muchas horas de entrenamiento, cuestión presentada también por ArcGIS y QGIS. Sin embargo, con la ayuda de los manuales que estas herramientas poseen, generalmente es posible usarlos sin mayor complicación. Además, existen cursos en línea, videos de YouTube e incluso puedes solicitar la guía de alguna Inteligencia Artificial para utilizarlos.

REFERENCIAS

Aguilar, M. L. y Pérez, M. (2014). *Estadística para los Negocios 2*. Universidad Panamericana.

Amemiya, T. (1981). Qualitative Response Models: A Survey. *Journal of Economic Literature*, *19*(4), 1483-1536. https://www.jstor.org/stable/2724565

Anderson, D. R., Sweeney, D. J., Williams, T. A. y García, G. S. (2016). *Métodos cuantitativos para los negocios* (11.ª ed.). Cengage Learning. http://latinoamerica.cengage.com

Anselin, L. (1988). Spatial Econometrics: Methods and Models. Kluwer Academic Publishers.

Anselin, L. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis*, *27*(2), 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin, L. (2023). GeoDa. https://geodacenter.github.io/

Aparicio, J. y Márquez, J. (2005). Diagnóstico y especificación de modelos panel en Stata 8.0. División de Estudios Políticos-Centro de Investigación y Docencia Económicas, México, 1-11.

ArcGIS. (2023). *Cómo funciona Autocorrelación espacial* (*I de Moran global).* https://desktop.arcgis.com/es/arcmap/latest/tools/spatial-statistics-toolbox/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm

ArcGis Pro. (2023). *Análisis de clúster y de valor atípico* (*I Anselin local de Moran)* (*Estadística espacial*). https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm

Arto, M. A. T. (2001). *El factor espacial en la convergencia de las regiones de la Unión Europea: 1980-1996*. Universidad Pontificia de Comillas.

Chasco, C. (2003). *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*. Dirección General de Economía y Planificación.

Chasco, C. (2006). Análisis estadístico de datos geográficos en geomarketing: el programa GeoDa. *Distribución y Consumo*, *16*(86), 34-47.

Chasco, C. (2012). *Manual fundamentos básicos de econometría* [Archivo PDF]. https://corochasco.files.wordpress.com/2017/02/fundamentos-bc3a1sicos-de-econometrc3ada_chasco_20122.pdf

Coronado Iruegas, A. A. (2021). *Geomarketing con Python.* https://abxda.medium.com/

Esparza Alba, D. (2023). *Machine Learning*. Apuntes Universidad Panamericana. [documento inédito].

Esri ArcGIS Pro. (2025). *What is ArcGIS*. https://www.esri.com/en-us/arcgis/geospatial-platform/overview

Greene, W. H. (2012). *Econometric analysis*. (7ª ed.). Prentice Hall.

Gujarati, D. (2011). *Econometrics by example*. Palgrave Macmillan.

Gujarati, D. y Porter, D. (2010a). *Econometría* (5.ª ed.). Mc. Graw Hill.

Gujarati, D. y Porter, D. (2010b). *Essentials of Econometrics* (4.ª ed.). McGraw Hill.

Hernández, J. (1995). *Introducción a la Econometría*. ESIC.

Instituto Nacional de Estadística y Geografía (INEGI). (2020). *Sistema Automatizado de Información Censal (SAIC). Censos Económicos 2019. Resultados Definitivos. Tabulados Interactivos*. https://www.inegi.org.mx/app/saic/default.html

Instituto Nacional de Estadística y Geografía (INEGI). (2025). *Mapa digital de México*. https://www.inegi.org.mx/temas/mapadigital/

Jarque, C. M. y Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, *55*(2), 163-172.

Kelejian, H. y Piras, G. (2017). *Spatial econometrics*. Academic Press.

Masini, J. y Vázquez, F. (2014). *Modelos cuantitativos de pronósticos* [Archivo PDF]. https://books.google.com.mx/books?id=fnLcBQAAQBAJ&printsec=frontcover&hl=es#v=onepage&q&f=false

Mayorga, M. y Muñoz, E. (2000). *La técnica de datos de panel una guía para su uso e interpretación*. Banco Central de Costa Rica. https://repositorioinvestigaciones.bccr.fi.cr/server/api/core/bitstreams/be8969f0-d0fd-4f45-8ff9-52b28cb9f64b/content

Molnar, C. (2021). *Aprendizaje Automático Interpretable: Una guía para hacer que los modelos de caja negra sean explicables*. https://fedefliguer.github.io/AAI/lineal.html

O'Halloran, S. (2005a). *Lecture 9: Logit/Probit* [Archivo PDF]. https://www.columbia.edu/~so33/SusDev/Lecture_9.pdf

O'Halloran, S. (2005b). *Lecture 10: Logistical Regression II—Multinomial Data* [Archivo PDF]. https://www.columbia.edu/~so33/SusDev/Lecture_10.pdf

Pratt, J. W. y Krasker, W. S. (1986). Bounding the Effects of Proxy Variables on Regression Coefficients. *Econometrica*, *54*(3), 641-656. https://doi.org/https://www.jstor.org/stable/1911312

Python. (2025). Python.

QGIS. (2025). QGIS. https://www.qgis.org/

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 31*(2), 350-371.

Rey, S. J., Anselin, L., Li, X., Pahle, R., Laura, J., Li, W. y Koschinsky, J. (2015). Open geospatial analytics with PySAL. *ISPRS International Journal of Geo-Information*, *4*(2), 815-836.

Seabold, S. y Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference.

Siebert, J., Groß, J. y Schroth, C. (2021a). A systematic review of python packages for time series analysis. ArXiv Preprint ArXiv:2104.07406.

Siebert, J., Groß, J. y Schroth, C. (2021b). linearmodels documentation (development version). https://bashtage.github.io/linearmodels/devel/

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica: Journal of the Econometric Society*, *48*(4), 817-838. https://www.jstor.org/stable/1912934

Wooldridge, J. M. (2015). *Introducción a la Econometría* (5.ª ed.). Cengage Learning.

ECONOMETRÍA
EN TU IDIOMA

HUGO BRISEÑO • DOLORES LUQUÍN

Se terminó de editar en noviembre de 2025

por Santi Ediciones (Rosario Ivonne Lara Alba)
Nance 1370, Colonia Del Fresno, Guadalajara, Jalisco.
www.santiediciones.com

UNIVERSIDAD
Panamericana

Nacido del aula y la inquietud estudiantil, *Econometría en tu idioma* transforma una disciplina compleja en una herramienta cercana, útil y reveladora. Con ejemplos prácticos, explicaciones detalladas y un enfoque pedagógico que respeta tanto la teoría como la intuición, los autores guían al lector desde los fundamentos de los modelos econométricos hasta su validación y aplicación en escenarios reales.

Ideal para estudiantes de finanzas, administración y economía, este texto no solo enseña a "hacer regresiones", sino a interpretar lo que los datos nos dicen sobre mercados, decisiones y fenómenos económicos. Una invitación a perderle el miedo a los modelos y a descubrir, en los números, las historias que moldean nuestro entorno.

UNIVERSIDAD Panamericana

UNIVERSIDAD Panamericana
Facultad de Ciencias Económicas y Empresariales

ISBN 978-607-8826-90-2