Miscelánea Matemática **81** (2025) 67-89 SMM

DOI: https://doi.org/10.47234/mm.8104

El impacto indeleble de la regresión

Carlos Vladimir Rodríguez Caballero
Departamentode Estadística
ITAM
vladimir.rodriguez@itam.mx
y
Daniel Ventosa-Santaulària
División de Economía
CIDE
daniel.ventosa@cide.edu

1. ¿Qué es una regresión?

El modelo de regresión constituye una de las herramientas estadísticas más socorridas para estudiar la relación entre variables. En su versión más sencilla, permite estimar la relación entre una variable usualmente llamada dependiente o de respuesta, y una o más variables llamadas independientes o explicativas. La variante más simple de este modelo es la denominada regresión lineal, que asume que la relación entre las variables puede describirse con una línea recta, lo cual se expresa como:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{1}$$

donde Y_i denota la i-ésima observación de la variable dependiente, X_i denota la i-ésima observación de la variable independiente, β_0 y β_1 son la ordenada en el origen (o intercepto) y la pendiente de la relación lineal simple que guardan X y Y. u es el término de error. Este último captura la variabilidad que el modelo «no puede» explicar. La estimación de los parámetros se realiza frecuentemente a través del método de **Mínimos Cuadrados Ordinarios**, MCO, que, como bien indica su nombre, minimiza la suma de las diferencias elevadas al cuadrado entre los valores observados y los valores predichos por el modelo.[1] La

Palabras clave: análisis de regresión, MCO, MV.
Este artículo forma parte de la celebración por los 50 años de la Licenciatura en Matemáticas Aplicadas en el Instituto Tecnológico Autónomo de México.

[1] En el apéndice de este artículo se provee un sucinto desarrollo del método MCO.

regresión proporciona una forma poderosa de entender y predecir comportamientos basados en datos históricos, permitiendo hacer inferencias sobre la relación entre variables y realizar pronósticos.

Un ejemplo de todo lo anterior podría resultar más elocuente y para ello aprovecharemos unas observaciones de alturas de padres e hijos recogidas por Sir Francis Galton hace ya más de cien años. Curiosamente, el término «regresión» fue acuñado tardíamente por Galton[2] a finales del siglo XIX, mientras investigaba justamente la relación entre las alturas de padres e hijos; se utilizó por primera vez en su obra para describir el fenómeno de que las alturas de los hijos de padres muy altos o bajos tienden a regresar hacia la altura media de la población general: el fenómeno de la regresión a la mediocridad.

Los datos originales de Galton se muestran en la figura 1,[3] en un diagrama de dispersión. Este es un vehículo muy cómodo para evidenciar relaciones entre variables. En el eje de las abscisas se mide la altura de los padres (nuestra X) mientras que en el de las ordenadas aparece la altura de los hijos correspondientes (nuestra Y). Si la interpretación de Galton es correcta, en nuestro diagrama se debería observar lo siguiente:

- Cuando los padres son «chaparros», los hijos suelen serlo menos.
- Cuando los padres son «altos», los hijos, también, suelen serlo menos.

Lo anterior efectivamente puede apreciarse en la figura 1. Lo interesante aquí es que el diagrama también permite suponer / hipotetizar que existe una relación entre la altura de los padres y la de sus hijos. Simplificando, se puede pensar que esa relación es lineal, como la de la ecuación (1).

Nosotros la estimamos usando MCO; está representada en la figura 1 con la línea roja. La ecuación estimada es: $Y_i = 60.81 + 0.65\ X_i$. Así, según este resultado, si el padre mide 164 centímetros, se podría

[2]Francis Galton fue un polímata inglés conocido por sus contribuciones en múltiples campos, entre los que destacan la estadística, la cartografía y la psicología. Primo de Charles Darwin, empezó estudiando medicina, pero abandonó esta para dedicarse a otras cosas, entre ellas, a las matemáticas. Galton tuvo su lado obscuro: fue uno de los primeros en promover el concepto de mejorar la raza humana mediante la selección de rasgos hereditarios; hasta le acuñó un nombre a tal práctica: eugenesia. Por increíble que parezca, Galton, en su afán de sustentar su propuesta creó los conceptos fundamentales de un primer curso de econometría o de modelos lineales: la correlación y la regresión a la media. El mismo hombre de ciencia también ha dejado para la posteridad su libro publicado en 1870 que lleva el título «*Hereditary genius: an inquiry into its laws and consequences*» [6], que es utilizado casi como un estandarte por los grupos supremacistas blancos en pleno Siglo XXI. Afortunadamente, la eugenesia ha sido debidamente vilipendiada y es hoy considerada una pseudo ciencia moralmente reprensible. Galton fue un hombre con claros y muchos oscuros, sin duda alguna.

[3]Datos: «Hereditary stature». Nature. 33 (848): 295-298. 1886b. Bibcode: 1886Natur..33..295.. doi:10.1038/033295c0

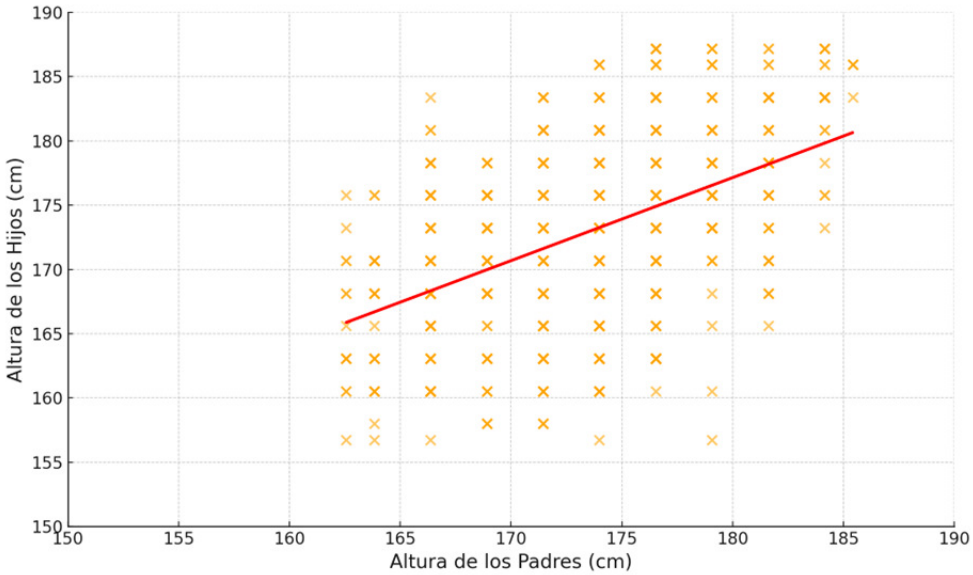

Figura 1. Diagrama de dispersión. Altura de padres e hijos. Fuente: Elaboración propia con datos de Galton (1886).

esperar que el hijo midiera 167.41 (menos chaparro), mientras que, si el padre mide 190 cm, el hijo debería salir de 184 cm (menos alto).[4]

Huelga decir que a este estudio le haría falta más profundidad. Si bien es razonable, por cuestiones genéticas, suponer que la altura de los padres está relacionada con la de sus hijos, sería importante considerar, además, muchos otros factores potencialmente relevantes, entre los que destacan, hábitos alimenticios, consumo de tabaco u otras drogas, ingesta de alcohol, condiciones medioambientales del entorno en el que viven las personas de la muestra, la profesión ejercida (minería, docencia, ...) y un largo etcétera al que biólogos, sociólogos y hasta economistas podrían contribuir. Ya con los controles adecuados (lo que implicaría estimar una regresión múltiple, es decir, con más de una variable independiente), podríamos muy posiblemente mejorar nuestra aproximación lineal de la relación. Lo anterior no es un tema trivial (es decir: qué sí incluir y qué no, en la regresión), por lo que se abordará al final de este artículo.

Permítanos, estimado lector, iniciar con una cuestión particularmente interesante: los orígenes del análisis de regresión.

[4]La «regresión a la mediocridad» también queda evidenciada por el hecho de que la pendiente es menor a 45 grados.

2. Orígenes controvertidos

El análisis de regresión tiene unos orígenes muy pintorescos. La potestad del método más conocido para estimar una regresión, MCO, ha sido objeto de una controversia que desnuda sin miramientos un lado poco atractivo de la ciencia, que es el ocasionalmente excesivo protagonismo de las personas que la llevan a cabo. El análisis de regresión fue revelado públicamente en un apéndice de apenas cinco hojas en el libro de Adrien Marie Legendre en 1805, «*Nouvelles méthodes pour la détermination des orbites des comètes*».[5] Legendre propuso MCO para ajustar los parámetros de las órbitas de los cometas, minimizando la suma de los cuadrados de las diferencias entre los valores observados y los valores calculados por la teoría, mejorando así la predicción de esos eventos astronómicos.

Si bien es cierto que Legendre fue el primero en publicar este método, Carl Friedrich Gauss afirmó poco tiempo después haberlo usado desde 1795, publicando su versión más extensa en 1809, en su libro «*Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*».[6] De ahí se siguió una controversia, principalmente de carácter epistolar, entre los autores así como con otros científicos que Gauss quiso usar como testigos de que él había ideado el método antes. La discusión no es particularmente edificante, pero sí muy entretenida.

Fue Gauss quien introdujo la noción de que los errores de observación se distribuyen normalmente y argumentó que el método de MCO era óptimo bajo la suposición de que los errores en las mediciones son independientes y normalmente distribuidos. Este enfoque conectó el modelo de regresión con teoría de probabilidad y sentó las bases de lo que luego se convertiría en la regresión lineal y la teoría estadística moderna. Aunque de Moivre y Laplace jugaron roles cruciales en el desarrollo temprano de la teoría detrás de la distribución Normal, fue Gauss quien realmente destacó su importancia y utilidad en el campo de la estadística y las ciencias aplicadas. Lo anterior, junto con sus publicaciones influyentes en el tema, contribuyó a que la distribución sea generalmente llamada «gaussiana» en honor, obviamente, a Gauss.

La distribución Normal fue descubierta por Abraham de Moivre inicialmente, un matemático francés que, en 1733 [2], derivó que la distribución binomial, conocida ya en esa época, se podía aproximar a una distribución Normal para grandes valores del tamaño de muestra, N. Más tarde, a finales del siglo XVIII, otro científico francés, Pierre-Simon Laplace expandió y generalizó el trabajo de De Moivre. Laplace

[5] Nuevos métodos para la determinación de las órbitas de los cometas [10].

[6] Teoría del movimiento de los cuerpos celestes que giran alrededor del sol en secciones cónicas [7].

desarrolló de forma más extensa las propiedades de la distribución Normal. Fue él quien, de hecho, utilizó esta distribución en contextos más generales y ayudó a establecerla como un concepto central en el campo de la estadística y la probabilidad. El teorema de De Moivre-Laplace es la versión más antigua de lo que hoy conocemos como el teorema del límite central.

El desarrollo formal de la técnica de regresión se consolidó con los trabajos de Karl Pearson y Sir Ronald A. Fisher, quienes ampliaron su uso para incluir múltiples variables, lo que es fundamental en la estadística moderna para el análisis de relaciones entre variables.

Podemos aprovechar el ejemplo de la relación de alturas entre padres e hijos para ilustrar la potencia y el alcance de las contribuciones de Gauss y Fisher al análisis de regresión. Considere una vez más el diagrama de dispersión y la línea de regresión mostrados en la figura 1. Recuerde que esa regresión se obtuvo mediante el cálculo de la línea (el intercepto y la pendiente, para ser precisos) que pasa más cerca de todos los puntos. Ello, como ya se explicó, se hizo minimizando la suma de las distancias al cuadrado de cada observación a la recta (observados menos estimados). Dado que el modelo de regresión así obtenido ya no es teórico, si no estimado, esas distancias ya no pueden ser denominadas errores y son usualmente referidas como residuales.

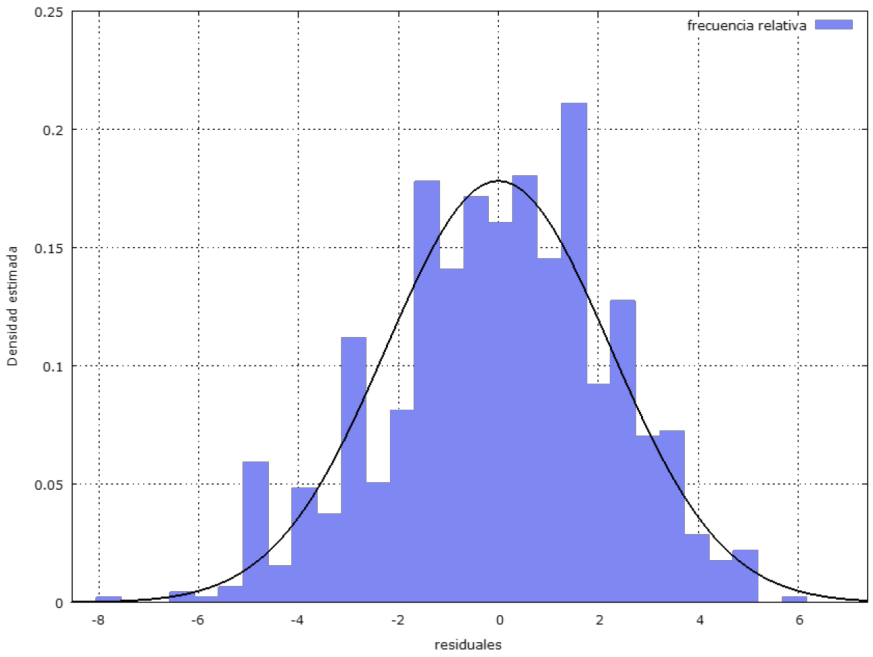

Figura 2. Distribución estimada de los residuales de una regresión. Fuente: Elaboración propia con datos de Galton (1886).

Observe en el histograma de la figura 2 cómo se distribuyen dichos residuales. Este ilustra con bríos que la propuesta de Gauss de asumir los errores (en el modelo teórico) y por ende los residuales (en su contraparte empírica) puede ser muy acertada. Una distribución Normal[7] está superpuesta en el histograma con objeto de reforzar el argumento.[8]

Este ejercicio permite también mostrar una de las grandes contribuciones de Fisher al análisis de regresión: la estimación por **máxima verosimilitud** (método conocido por sus siglas, MV). El método de MCO, hay que insistir, no es la única forma de estimar un modelo de regresión, es solo una magnifica opción para hacerlo. Pero igualmente valiosa es MV.

El método de MV fue desarrollado en 1922 por Ronald A. Fisher,[9] uno de los estadísticos más influyentes del siglo XX.[10] La técnica se basa en la selección de los valores de los parámetros de un modelo estadístico que maximizan la función de verosimilitud, es decir, que hacen que los datos observados sean los más probables bajo el modelo especificado. Esta forma de estimar un modelo de regresión es un referente de la inferencia estadística y se emplea en cantidad de disciplinas científicas.

El método de MV está apoyado por la siguiente propuesta. En vez de minimizar la distancia cuadrática entre las observaciones y el modelo para estimar β_0 y β_1 (como en MCO), MV se enfoca, en el contexto de un modelo de regresión, en ajustar los parámetros del modelo de tal suerte que los residuales (otra vez: la diferencia entre los valores observados y los valores predichos por el modelo) se comporten como realizaciones de una distribución específica; frecuentemente (pero no siempre necesariamente) se asume que dicha distribución es normal, con media cero y varianza constante.

La función de verosimilitud (que no es sino la función de densidad en la que aparecen explícitos los parámetros a estimar) se maximiza para encontrar los coeficientes del modelo (pendientes e intersección, en el caso de una regresión lineal) que hacen que la observación de los datos sea la más probable, asumiendo que los residuales siguen esta distribución normal. Note que, siguiendo el ejemplo con una normal, tendremos que estimar la esperanza (dónde aparecen los parámetros)

---

[7] De esperanza cero ($1 \times 10 - 12$, para ser exactos) y varianza 2.24.

[8] Considere que, si la distribución Normal no resultara tan adecuada, existen muchas otras opciones. Por ejemplo, si las colas de la densidad fueran notoriamente más pesadas, quizá convendría una t de student.

[9] Fisher introdujo este método en un artículo titulado «*On the mathematical foundations of theoretical statistics*», [4]. Cabe resaltar que en el apéndice también se proporciona un escueto desarrollo del método.

[10] Vale la pena resaltar que Fisher tenía el sucio hábito de fumar y nunca aceptó la evidencia de que el tabaquismo era el origen de cantidad de enfermedades... *En casa del herrero, azadón de palo...*

y también la varianza (que es lo que se precisa para definir la distribución). Sus desventajas con respecto a MCO son pocas, pero no triviales. MV precisa que se especifique la distribución (de los residuales, si se aplica al modelo de regresión). Si se escoge la distribución inadecuada, la inferencia podría no ser muy buena.[11]

No obstante, MV ofrece ventajas innegables con respecto a MCO. En primera instancia, toma en cuenta la teoría de la probabilidad de manera más natural en el proceso de estimación; MCO es un método eminentemente geométrico (minimización de distancias). En segunda instancia, permite, también de manera muy natural, encontrar distribuciones de los parámetros que son necesarias para realizar inferencia estadística sin necesidad del uso avanzado de teoremas de convergencia, y por otro lado, realizar estimaciones de modelos de regresión más complejos (no lineales, por ejemplo). Sin ahondar más en tecnicismos innecesarios en este artículo, podríamos resumir la relevancia de los estimadores de MV de la forma siguiente. Si nuestro supuesto distribucional es acertado, entonces nuestras estimaciones vía MV serán «mejores» que aquellas vía MCO.[12] La palabra «mejor» es de absoluta relevancia en teoría estadística, solo que la disfrazamos con una entidad abstracta más simpática: le llamamos «eficiencia» a lo relacionado estrictamente con la varianza; recuerde que el estimador también es una variable aleatoria, por lo que tiene esperanza, varianza (y, para generalizar, función de distribución). Así, cuando un estimador es más eficiente que otro, ello implica que tiene menor varianza, y por ende las realizaciones suelen caer más cerca de la esperanza.

Hay más alternativas a MCO; está, por ejemplo, el método de momentos, que fue inventado por los matemáticos Pafnuti L. Chebyshev y Karl Pearson en 1887 [1] y 1894 [11], respectivamente.[13] Pearson desarrolló este método estadístico como una técnica para estimar los parámetros en distribuciones de probabilidad, basándose en los momentos muestrales. El procedimiento implica igualar los momentos teóricos derivados de la distribución de probabilidad con los momentos muestrales correspondientes, proporcionando así una forma de estimar los

[11] Existen ya alternativas, como el estimador de cuasi-máxima verosimilitud o por ejemplo los estimadores no-paramétricos.

[12] Note, no obstante que, si la especificación es lineal, los regresores no son estocásticos y los errores se distribuyen como una Normal, entonces el campeón es MCO, y ostenta en ese caso el título de Mejor Estimador Lineal Insesgado.

[13] El origen formal del método de momentos puede rastrearse hasta Pafnuti Chebyshev. Él utilizó los momentos en sus estudios de probabilidad y teoría de la estadística, sentando las bases teóricas del método. Sin embargo, fue Karl Pearson quien popularizó y empleó el método de momentos en estadística aplicada y lo introdujo en su forma moderna, especialmente en el contexto de la estimación de parámetros de distribuciones. Como nota miscelánea, sepa que originalmente, Pearson de llamaba Carl, pero se hizo cambiar el nombre a Karl en honor a... Marx. Como uno de los autores de este ensayo cuyos padres no llegaron a tanto (ni él mismo por cierto) y qué bueno, caray.

parámetros. La versión generalizada, propuesta por el econometrista Lars Peter Hansen en 1982 [9], implica hacer más igualaciones de momentos muestrales con teóricos que parámetros a estimar. Ello implica resolver un sistema sobre identificado lo cual no es, sorprendentemente, malo. Esta técnica es muy popular en econometría.

Al margen de la técnica de estimación, lo sustantivo es que el análisis de regresión se emplea en una miríada de disciplinas; destacan, la investigación médica (especialmente la clínica) y las ciencias de la salud en general, la psicología, la ingeniería (especialmente en control de calidad), la biología (particularmente las ciencias ambientales), la física, las ciencias sociales en general y la economía en particular. En este último caso, la estadística y la economía se fusionaron creando lo que hoy se conoce como econometría. Esta rama tan especial de la estadística exige conocimientos en economía en el diseño e interpretación de los modelos y se preocupa desmesuradamente por los problemas relativos a estudios en los que los datos no provienen de un experimento controlado, lo que resulta en un sinfín de problemas a veces muy difíciles de resolver.[14]

3. ¡La regresión es una media!

La idea de la regresión es simple y sencillamente genial. Para entenderla basta conocer unos pocos conceptos matemáticos. Debemos empezar por algo fundamental, el concepto de «aleatoriedad». Este se refiere a eventos o fenómenos que ocurren de manera impredecible y sin un patrón determinable, siendo el resultado de factores que no siguen una

[14]La Comisión Cowles, fundada en 1932 por Alfred Cowles y actualmente mejor conocida como *Cowles Foundation for Research in Economics* de la Universidad de Yale, tuvo un papel importante en el desarrollo y la formalización de la econometría moderna, incluyendo las técnicas de estimación estadística empleadas en modelos de regresión. Promovió el uso de métodos estadísticos rigurosos en la teoría económica, y, en voz de importantes economistas como Trygve Haavelmo clarificó la importancia de los fundamentos probabilísticos en los modelos económicos. La contribución al desarrollo de la teoría de la identificación es enorme. Para el caso que nos ocupa, la Comisión Cowles fue instrumental en promover el uso de MCO y de Máxima Verosimilitud en la econometría. El economista noruego Ragnar Frisch fue quien acuñó el término «*Econometrics*» [5]. Ganador del primer Premio Nobel de Ciencias Económicas en 1969, fue uno de los fundadores de la prestigiosa *Econometric Society* (nacida en periodo entre guerras con el objetivo de facilitar la comunicación entre economistas estadounidenses y europeos) y primer editor de la revista Econometrica. Palabras más, palabras menos, Ragnar escribió en el primer volumen de Econometrica en 1933 «La econometría es la unificación de la teoría estadística, la teoría económica y la teoría matemática. Cada una de estas áreas, es por sí misma necesaria, pero no suficiente para un adecuado entendimiento de las relaciones cuantitativas en la vida económica moderna.» Frisch fue supervisor doctoral del noruego Trygve Magnus Haavelmo, quien también ganaría el Nobel de economía en 1989 por haber clarificado los fundamentos de la teoría econométrica. La econometría ha cambiado mucho en el último siglo. A los autores les gusta la siguiente definición de econometría puesto que tiene una visión más global y moderna: «En términos generales, la econometría tiene como objetivo brindar un contenido empírico a las relaciones económicas para probar teorías económicas, realización de pronóstico, toma de decisiones y para evaluación de políticas.» De Geweke, J., J. Horowitz, y M.H. Pesaran de 2008 [8].

regularidad específica. Algo aleatorio implica que cada resultado posible tiene una probabilidad conocida, pero no se puede predecir con certeza cuál ocurrirá en una instancia particular. Para domarla, emplearemos otra herramienta, a la que llamaremos «variable». En particular necesitamos una que sea, no lo van a creer, aleatoria. La variable aleatoria (v.a.) es una función que asigna un valor numérico a cada posible resultado de un experimento aleatorio. Por ejemplo, al tirar un dado, esos valores posibles son, naturalmente, 1, 2, 3, 4, 5, y 6. Aunque el resultado es incierto, no todo está perdido: a cada posible resultado, se le puede asignar una probabilidad. Se trata de una medida de cuán probable es que ocurra un determinado evento y se expresa mediante un número que debe estar cernido entre cero y uno, donde cero significa que la probabilidad de que ocurra el evento es nula y uno significa que el evento ocurrirá con absoluta certeza. En el dado, si no está cargado, la probabilidad es la misma para todos los posibles resultados: $\frac{1}{6} \approx 0.16666$. Lo anterior puede expresarse de manera más elocuente con la figura 3.

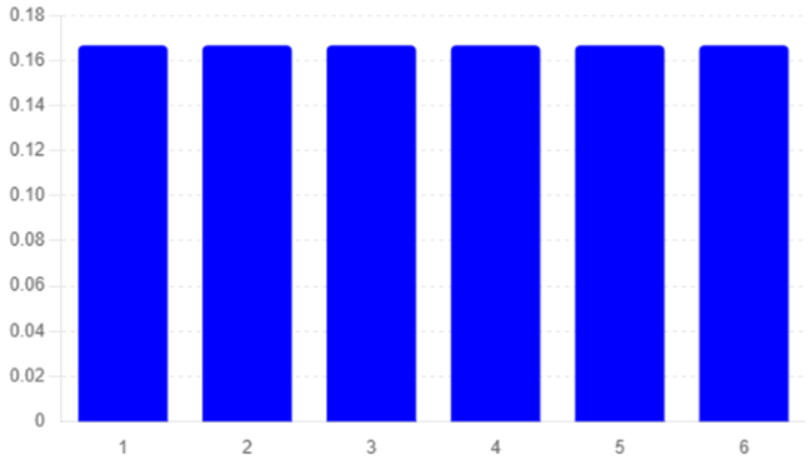

Figura 3. Función de distribución de una v.a. uniforme discreta, idónea para describir el comportamiento incierto de un dado no cargado. Fuente: Elaboración propia.

Esta figura nos tipifica qué tan probable es que caiga un lado específico del lado: ofrece un panorama muy claro de lo que puede ocurrir en el experimento. La función matemática que nos permite trazar la figura no es otra cosa si no la función de probabilidad de la variable aleatoria. Toda esta información es útil, ya que reduce la incertidumbre, especialmente, la de los ludópatas. Hablando ya más en serio, el concepto de variable aleatoria (y de su distribución de probabilidad asociada) lo usamos para estudiar fenómenos cuyos resultados desconocemos de forma precisa. Por ejemplo, el desempeño escolar, el resultado de una elección, el cálculo de un costo de seguro de vida, y un muy largo etcétera.

No siempre es fácil contar con la distribución del fenómeno aleatorio; a veces, de hecho, es virtualmente imposible ex ante, especialmente si aplicamos el concepto a fenómenos que están fuera del ambiente controlado de un experimento (considere, por ejemplo, los resultados de una elección). Pero usualmente podemos consolarnos un poco con alguna característica específica de la distribución. Es, para empezar, particularmente útil tener una idea de dónde está ubicada la distribución, es decir, tener una medida de localización, misma que nos permite saber alrededor de qué valor puede revolotear el resultado de un experimento. También es muy útil saber qué tan lejos de ese centro puede caer el resultado, eso es una medida de dispersión. Ambas medidas son muy socorridas; la más conocida (aunque a veces mal entendida) es la primera, y le llamamos «media». La segunda es la «varianza». Concentrémonos en la primera. Como justamente comentamos arriba, la media brinda la información de por donde andan los posibles resultados. Por eso la usamos, entre muchas otras cosas, para evaluar el desempeño de un estudiante; si su promedio es, digamos, ocho, ello no quiere decir que el estudiante en cuestión vaya a sacar ocho siempre; más bien implica que su calificación estará alrededor de ese número. Dicho de manera cruda y cruel, la media en sí puede ser usada como una medida de predicción. ¿Qué tan buena será? De eso nos encargaremos más adelante. Ahora bien, si queremos saber qué tan lejos de ocho podrían caer las calificaciones del estudiante, habremos de recurrir a la varianza. Si nos importa saber si es más probable que la calificación sea menor de ocho a que sea mayor de ocho, entonces habremos de medir el sesgo de la distribución. Si lo que nos preocupa radica en los eventos extremos (sacar cero o diez), entonces mediremos la curtosis de la distribución. Estos cuatro conceptos están emparentados. Se trata de distintos momentos de la distribución. Más precisamente, los primeros cuatro momentos nos ayudarán a encontrar estas características distribucionales.[15]

Para calcular el valor medio o esperado del experimento basta con tomar cada posible valor que pueda adoptar la variable aleatoria (cada posible resultado) y multiplicarlo por la probabilidad de que ocurra. En el caso del dado, la fórmula es sencilla:

$$\boxdot \times \frac{1}{6} + \boxdot \times \frac{1}{6} + \boxdot \times \frac{1}{6} + \boxdot \times \frac{1}{6} + \boxdot \times \frac{1}{6} + \boxdot \times \frac{1}{6} = 3.5.$$

Note que nadie tiene la expectativa de obtener 3.5 en el tiro de un dado, aunque si el dado se queda en la orilla de la pared medio volteado

[15]La media es el primer momento, la varianza tiene que ver con el segundo y, a diferencia de la media, está centrada (se le resta la media), por lo que le llamamos segundo momento central. El sesgo y la curtosis son el tercer y cuarto momentos estandarizados. También están centrados, pero, además, se dividen por la desviación estándar, que es la raíz cuadrada de la varianza. Vale la pena aclarar que, sorprendentemente, algunas distribuciones carecen de dichos momentos; eso se estudia en un primer curso de probabilidad.

entre 3 y 4 no faltará quien alegue lo contrario. Ya en serio, dicho valor indica un centro alrededor del cual están los posibles resultados que puede adoptar el experimento de haber lanzado el dado. Dicho cálculo sencillo se le denomina Esperanza Matemática en teoría de la probabilidad y se denota con una (poco imaginativa) letra E. Si a la v.a. la llamamos Y, entonces lo que hemos calculado es:

$$E(Y) = \sum_{i=1}^{6} Y_i \; P(Y = Y_i).$$

En la práctica, no siempre conocemos las probabilidades (de hecho, casi nunca), pero, por suerte, podemos tirar muchas veces el dado, 100, 200, 1000 veces (lo que corresponde a una muestra). Aunque no idéntico, el número de veces que salga cada posible valor del dado será parecido a $\frac{1}{6}$, especialmente si la muestra es grande. De hecho, entre más experimentos hagamos, más parecidas a $\frac{1}{6}$ se irán haciendo las frecuencias y más se irá aproximando la media muestral a la esperanza 3.5. Lo anterior es uno de los resultados más poderosos de la estadística, la ley de grandes números en su concepción más simple.[16]

Ahora se requiere extender el concepto de esperanza, permitiendo incorporar más información al cálculo. Considere el siguiente ejemplo: obtener el promedio de calificaciones de un alumno puede ser una medida práctica para estimar su posible desempeño en el semestre en curso. Esa información es muy útil en sí, pero la podemos robustecer. ¿Qué tal si añadimos información pertinente del estudiante? Por ejemplo, podríamos calcular su promedio de acuerdo con el número de horas que estudia, o en función de la distancia a la que vive de la escuela, o del nivel de estudios de los padres, o de su nivel socio económico o inclusive de su género (esperen, no desenvainen sus espadas, líneas abajo ahondaremos en este respecto). Contar con toda esta información nos permitirá saber alrededor de qué valor está el centro de la distribución condicionado a que haya estudiado mucho o poco, a qué viva cerca o lejos, a que venga de un hogar donde los padres tienen muchos estudios o no los tienen, etcétera. Esa información es, genuinamente, más útil y valiosa que la del simple promedio, ¿no cree? Lo es porque permite plantear preguntas interesantes a los datos y eventualmente ofrecer respuestas sólidas. Alguien con interés en temas de equidad, puede estudiar si el desempeño escolar depende de la situación económica de sus progenitores. Si el promedio de calificaciones entre estudiantes de hogares más precarios y estudiantes de hogares más holgados es (estadísticamente) igual, entonces habrá encontrado su respuesta: no, el nivel de

[16] El otro gran resultado de la estadística es el teorema del límite central mencionado anteriormente. Este nos señala que, bajo condiciones bastante generales, cuando estimamos una media, su comportamiento aleatorio cada vez es más parecido al de una variable aleatoria gaussiana.

ingreso familiar no parece estar relacionado con el desempeño escolar. Lo mismo puede plantearse en términos de género. Si hay diferencias estadísticas significativas entre los promedios de las y los estudiantes, entonces, sabiendo que la capacidad cognitiva es igual entre hombres y mujeres, lo que se habrá obtenido es evidencia de discriminación por género. Y sí, hay investigaciones que parecen apuntar hacia allá. Hay algo muy interesante que se cuece en el párrafo anterior. ¿Qué información pertinente debemos añadir? Bueno, esa respuesta no es sencilla. Es el trabajo de muchas áreas de estudio diferentes en torno a una misma pregunta. En el ejemplo sencillo de los factores que pudieran intervenir en las calificaciones del estudiante, habría que escuchar a otras áreas científicas, como la pedagogía, las ciencias de la salud, tal vez hasta la antropología. ¿Se imaginan la cantidad de trabajo multidisciplinario que se necesita para calcular de la mejor forma posible nuestro simple promedio?

La esperanza sujeta a que se cumplan ciertas condiciones se denomina, no muy originalmente, esperanza condicional. Pues aquí viene la gran revelación. La esperanza condicional no es otra cosa sino el dichoso modelo de regresión de Legendre y Gauss. Y sí, también se puede estimar con datos, igual que como estimamos la esperanza usando un promedio muestral simple. Ahondemos al respecto.

Suponga que nosotros codificamos el ingreso de los egresados de una universidad en una v.a., a la que llamamos Y. Sospechamos que la distribución del ingreso de los egresados puede cambiar en función de otras variables, posiblemente aleatorias también (aunque no necesariamente) que aglutinaremos en X. Entre esas variables, pueden estar: el desempeño escolar, la situación del país (al momento de egresar), la condición económica de los padres y de las madres, etcétera. Pues bien, con base en lo anterior, planteamos la esperanza del ingreso del estudiante i que egresa condicionado a todos esos factores, $E(Y_i|X_i)$, y la comparamos con la variable aleatoria misma: $Y_i - E(Y_i|X_i)$. Como ya explicamos, una esperanza no da el valor que va a adoptar la variable en sí, si no la localización del centro de su distribución. Por ello, resulta obvio que $Y_i - E(Y_i|X_i)$ difícilmente será igual a cero. Pongámosle una letra a dicha diferencia para identificarla matemáticamente; usamos la letra u_i. Juguemos con la ecuación:

$$\begin{aligned} Y_i - E(Y_i|X_i) &= u_i, \\ Y_i &= E(Y_i|X_i) + u_i. \end{aligned}$$

Pues bien, la línea de regresión poblacional $E(Y|X)$, que algunos prefieren notar como $m[X]$ más el término de error u nos dará el llamado

modelo de regresión. Lo más importante a retener es que una regresión no es otra cosa si no un promedio, condicionado (y a veces muy sofisticado si ustedes quieren), pero un promedio, al fin y al cabo.[17]

Ahora bien, esa esperanza condicional aún suena muy etérea, vaga. Habrá que ir pensando en definirla con más precisión. La verdad es que podemos pensar en diversas ecuaciones matemáticas para $m(X)$, aunque, de hecho, siempre se comienza con la opción más simple entre todas las posibles elucubraciones, la ecuación de una recta (como ya vimos con el primer ejemplo, una línea con intercepto y una pendiente constante). Especifiquemos pues la esperanza condicional como una relación lineal entre Y y X para que el concepto se vuelva más tangible, es decir, usemos $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$. Notará que, como lo habíamos señalado, la esperanza de Y_i varía en función del valor que adopte X_i. Recuerde que en nuestro ejemplo Y_i es el ingreso del egresado; suponga que solo usamos una variable X_i, que es el ingreso de sus progenitores. Si al estimar β_1, este nos sale estadísticamente igual a cero, habremos encontrado evidencia de que el ingreso familiar no afecta el ingreso del egresado o la egresada. Este ejemplo ilustra el enorme potencial científico de un modelo de regresión.

4. Estimando medias condicionales

4.1 De los cometas al impacto de estudiar una licenciatura

Ahora sí, ya que sabemos que una regresión es una esperanza condicional y tenemos indicios de como especificarla, habrá que preocuparse por estimarla. Cuando se trata de una media simple, el mejor estimador es la media muestral, consistente en sumar los valores de las realizaciones y dividir dicha suma por el número total de realizaciones, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, donde N es el tamaño de la muestra. Pues bien, para calcular una media condicional, tenemos un coctel de opciones. La más popular, por muchas razones, es la técnica inventada por Legendre en 1805 (aun cuando Gauss haya insistido haberla desarrollado en 1795 dominado por el soponcio de publicarlo): MCO (u OLS, por sus siglas

[17] Vale la pena mencionar que, sin que sea necesario precisar las propiedades de las variables Y y X, es posible conocer propiedades en extremo interesantes de u: la esperanza, condicional o incondicional de u es cero; No existe relación lineal alguna entre u y cualquier transformación de X. Estos resultados se logran mediante otra herramienta genial, que es la Ley de Esperanzas Iteradas. Esta establece que la esperanza de una variable aleatoria se puede calcular iterativamente a través de una expectativa condicional: $E(Y) = E[E(Y|X)]$. Con base en esta mismita ley se puede demostrar que la varianza de Y es la suma de la varianza del modelo de regresión (lo que se «puede» explicar) y la varianza del término u (lo que no se «puede» explicar). Es un resultado fantástico que permite tener una idea de la capacidad explicativa de un modelo de regresión: es el análisis de varianza, o ANOVA, como se suele denominar.

en inglés). Retomando el ejemplo anterior, el parámetro que nos interesa estimar, β_1, se puede obtener, usando MCO, mediante la sencillísima fórmula (ver el apéndice):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

4.2 Cuidado con los sinsentidos

Aquí es importante entender que puede (y en general debe) haber más regresores. ¿Recuerda el comentario, en el primer ejemplo, que señala que lo más difícil al realizar un análisis de regresión es decidir las variables que conviene incluir y las que conviene excluir? Si estimamos la relación que guarda el ingreso del egresado con el ingreso familiar, es importante asegurarnos en la medida de lo posible que dicha relación no sea espuria, en el sentido que sea otra la fuente que influya tanto a la variable Y como a la X, mientras que X y Y no tengan relación alguna. Para entender el riesgo, vale la pena mostrar varios ejemplos:

- Se ha observado que cuando las ventas de helados aumentan, la tasa de mortalidad en personas de edad avanzada tiende a aumentar también.
- Un famoso estadístico, Udny Yule relacionó en 1926 la tasa de matrimonios (celebrados por la iglesia anglicana) y la tasa de homicidios en Inglaterra y Gales; encontró una correlación positiva y significativa entre matrimonios y homicidios.[12] Los más cínicos quizá pensarán que ahí «hay algo».

En ambos ejemplos, la aparente relación se debe a una tercera variable que afecta a las otras dos, pero no está incluida en la especificación del modelo: en el caso de la relación entre las ventas de helados y las muertes de personas de edad avanzada, la variable omitida es la temperatura: una mayor temperatura insta a la gente a consumir más helado, y al mismo tiempo (y de manera independiente) afecta desproporcionadamente la salud de la gente mayor. En el ejemplo de la relación entre matrimonios y homicidios, la variable omitida es algo más sutil, una tendencia secular: los datos de Yule corresponden a un periodo en el que simultáneamente, (i) el fervor religioso declinaba y hacía que menos gente se casara usando un rito religioso, y, (ii) la sociedad se volvía menos violenta. Ello se tradujo en menos muertes por asesinato. En otras palabras, faltó controlar por el tiempo, mismo que refleja los cambios sociales que afectan las prácticas de los individuos en múltiples dimensiones.

Los estadísticos y econometristas deben cuidarse de la regresión espuria. Tanto si es un experimento controlado como si se trata de datos

económicos, el investigador debe ponderar a conciencia lo que incluye en su modelo de regresión. Si peca por omisión, corre el riesgo de confundir relaciones espurias con genuinas. Tristemente, tampoco conviene pecar por exceso, cuestión que por cierto también tiene nombre: «sobreajuste». El problema del sobreajuste (*overfitting*, en inglés) en el marco del análisis de regresión ocurre cuando un modelo estadístico describe «demasiado bien» los datos disponibles, capturando no solo las relaciones subyacentes reales sino también el ruido y las fluctuaciones aleatorias presentes en los datos. Esto puede llevar a un desempeño deficiente del modelo cuando se aplica a nuevos datos no observados (otras muestras o pronósticos, por ejemplo), ya que el modelo se ha ajustado demasiado a las particularidades del conjunto de datos original y no ha sido capaz de capturar eficazmente la relación general, que es la que realmente importa.

En diversas áreas del conocimiento, como pueden ser las ciencias económicas, las políticas, o las ambientales, los investigadores se encuentran una y otra vez con el problema de tener muchos datos y una teoría que no ha alcanzado los niveles óptimos de madurez. Esto conlleva la penosa realidad de que es ciertamente muy fácil llegar a confundir el ruido con la señal, Es una forma natural de entender el sobreajuste en estadística. ¿Cómo impacta esto a nuestro modelo de regresión? Veámoslo con un poco más de detalle. En el modelo de regresión, la parte correspondiente a la esperanza condicional representa la señal (*signal*, en inglés) mientras que el término de error corresponde al ruido (*noise*, en inglés), es decir:

$$Y = \underbrace{E(Y|X)}_{\text{señal}} + \underbrace{u}_{\text{ruido}}.$$

El Santo Grial estadístico es «identificar» (otra palabreja técnica adorada por los especialistas) la señal. Cuando ello ocurre, se traduce en realizar un buen ajuste. Cuando el ajuste es paupérrimo, es decir cuando no hemos logrado capturar la señal emitida por $E(Y|X)$, se dice que tenemos un subajuste (*underfitting*, en inglés). En cambio, cuando exageramos un poquitín y logramos inclusive ajustar el ruidoemitido por u más allá de solo descubir la estructura subyacente de los datos, entonces estamos sobreajustando nuestros datos (*overfitting*, en inglés). Un ejemplo hipotético ayudará a explicar el concepto. Suponga que dispone de dos variables climáticas, X y Y. Suponga además que la esperanza condicional que nos interesa, de Y, viene dada la siguiente ecuación cúbica:[18]

$$Y_i = a + b\, X_i + c\, X_i^2 + d\, X_i^3 + u_i.$$

[18]Destaca que la ecuación es lineal en los parámetros, por lo que se podría estimar con MCO.

Como ya hemos explicado, lo que se busca es estimar la esperanza condicional dada por la regresión poblacional cúbica. Esta se representa con la línea negra continua en la figura 4. No obstante, imaginemos que nuestro conocimiento sobre la ecuación climática está algo más que incompleto (básicamente en pañales), en un grado superlativo, e, ingenuamente consideramos que la relación entre Y y X es lineal, $E(Y_i|X_i) = a + b\ X_i$. Esto conlleva al subajuste mostrado con la línea verde punteada (también en la figura 4). Ahora bien, en nuestro afán de «hacer ciencia» como si se tratase de dibujos de unión de puntos tan divertidos en la infancia, alguien sugiere un método que logra conectar diversos puntos de forma *maquiavélica*; la línea punteada roja ilustra esta situación en la que sobreajustamos la relación climática, es decir: el modelo de regresión no solo ajusta la relación cúbica, que es lo que nos interesa, sino también las desviaciones aleatorias de las observaciones. Y aunque el ajuste se ve espectacular, el sobreajuste termina obscureciendo nuestro ya de por sí escaso conocimiento de la relación entre las variables Y y X En otras palabras, ya no estamos estimando bien la media condicional de Y, que era el objetivo. En este sentido, vale la pena recordar las sabias palabras del gran matemático John von Neumann:[19] «Con cuatro parámetros puedo ajustar un elefante y con cinco puedo hacer que menee su trompa».[20]

Hasta aquí debe quedar claro que en estadística (y en econometría) el interés primordial es capturar correctamente $E(Y|X)$. Para ello debemos mantener un balance entre las varianzas de la señal y del ruido; a eso le llamamos la relación señal-ruido, o más habitualmente, el *signal-to-noise ratio* en el argot técnico. Cuando la varianza de u es mayor a la de X diremos que el ruido domina a la señal débil, en caso contrario diremos que la señal es más potente que el ruido por lo que la domina. El segundo caso, huelga decir, es el ideal. En la práctica, y usando la ocasionalmente elegante jerga técnica, lo difícil es identificar la señal sin ahogarse en el ruido.

Estimar correctamente un simple promedio (condicional) es una tarea más compleja de lo que podría parecer en un principio. Buena parte del esfuerzo de un estadístico y/o un econometrista radica en especificar exitosamente la esperanza condicional. Aprender a hacer esto precisa, infortunadamente, cursar varias materias formativas en las áreas de estadística y econometría. Es todo un largo camino a la media que recorrer, pero vale mucho la pena.

[19] Un brillante científico estadounidense de origen húngaro.
[20] Atribuida a Neumann en Dyson, F. (2004) [3].

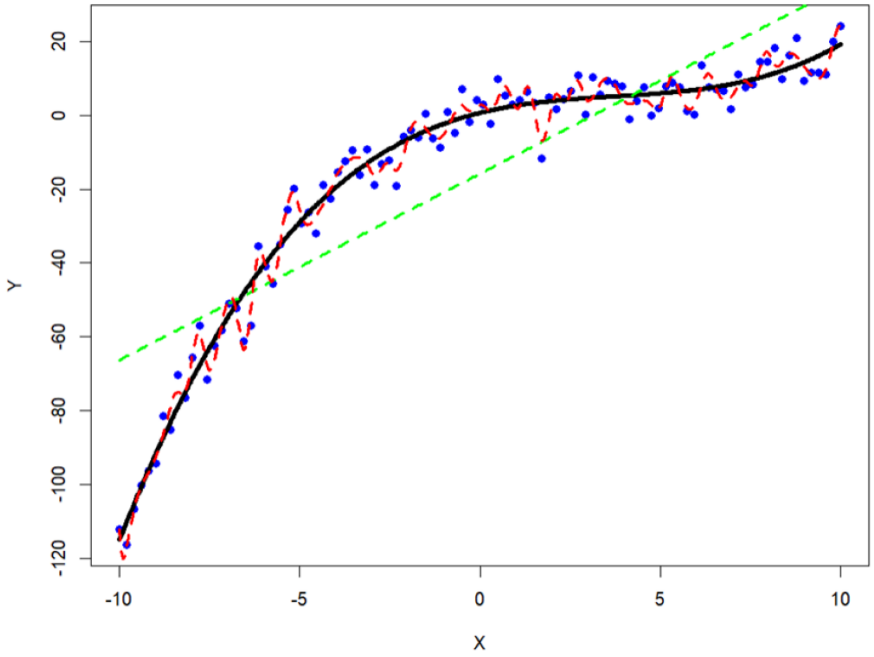

Figura 4. ajuste, subajuste y sobreajuste de un modelo. La relación real es cúbica. La línea negra representa la relación real entre variables y es una ecuación cúbica. La línea verde punteada es la representación de un subajuste y la roja de un sobreajuste. Fuente: Elaboración propia.

5. ¿Estudiar estadística o no estudiarla? En promedio siempre le convendrá

El análisis de regresión ha dejado una marca indeleble en numerosas disciplinas, consolidándose como una herramienta fundamental para entender y modelar relaciones entre variables. Desde su origen en la astronomía, cuando fue utilizado por Legendre y Gauss para ajustar las órbitas de los cometas, hasta su adopción en la economía, la biología, la ingeniería y las ciencias sociales, este método ha evolucionado para abordar problemas cada vez más complejos. Su capacidad para desentrañar patrones ocultos en datos, prever comportamientos futuros y validar hipótesis científicas ha revolucionado la manera en que investigadores y profesionales abordan el análisis cuantitativo. El análisis de regresión no solo permite hacer predicciones útiles, sino que también ofrece una base sólida para la toma de decisiones informadas en contextos donde la incertidumbre es la norma. A través de su desarrollo, desde los métodos clásicos de mínimos cuadrados hasta las técnicas estadísticas y econométricas más avanzadas, la regresión ha demostrado ser una herramienta versátil y poderosa, cuyo impacto perdura tanto en la investigación como en la práctica profesional diaria.

En este artículo hemos presentado el modelo de regresión, junto con la triste historia de la técnica más famosa para estimarlo (MCO); hemos sentado también las bases de su construcción (la esperanza condicional) así como el alcance y las limitaciones que puede llegar a tener esta importantísima herramienta científica. Tema fundamental en cualquier curso de estadística y econometría, es a veces temida y repudiada por el estudiantado por ser considerada una materia obscura, casi tanto como esa que emplean los físicos para explicar todo lo que no entienden del universo. Y es cierto, luego los investigadores no tienen empacho con su terminología ostentosa, ni les importa hablar de multicolinealidad, homoscedasticidad (y su contrario, heteroscedasticidad), endogeneidad, independencia (y no de un tirano), autocorrelación, heterogeneidad, no linealidad, identificación, errores gaussianos, choques estructurales, causalidad, granger-causalidad, granger causalidad a la Breitung-Candelon, micronumerosidad, simultaneidad, errores de medición, sesgo, eficiencia, ergodicidad, estacionariedad, alopecia,[21]... en fin. Pero no se deje apantallar, lo que a usted le están enseñando es a calcular promedios correctamente.

Apéndice: El modelo de regresión y su resolución vía MCO y MV

Para explicar la estimación del modelo, aunque sea de manera introductoria, vale la pena primero entender gráficamente qué es lo que vamos a estar haciendo y por qué es necesario el uso del cálculo diferencial en esta etapa.

Los círculos azules en la figura 5 muestran los datos observados de dos variables X, y Y. Suponga que queremos encontrar una ecuación que nos sirva para predecir la observación Y_i dado que sabemos el valor que tiene la observación X_i. La recta (línea en negro) es la función de regresión poblacional. Es teórica, es decir que no la vamos a observar, desafortunadamente; solo vive en nuestra imaginación matemática. La función de regresión poblacional justamente es la ecuación con la que definimos nuestra esperanza condicional que tanto hemos mencionado y, a partir de la cual, podremos a la postre, hacer nuestras dichosas predicciones. El círculo rojo cuyo valor es $E(Y_i|X_i)$ representa nuestra predicción del valor observado Y_i. Ahora bien, es evidente que en nuestro caso la predicción (círculo rojo) no es igual al valor observado (círculo azul). Por ende, estamos generando un error de predicción, el cual es común denotarlo como u_i que sería igual a la distancia entre Y_i y $E(Y_i|X_i)$. Así obtenemos la ecuación $Y_i = E(Y_i|X_i) + u_i$.

[21] Este último problema preocupa seriamente a los autores del ensayo.

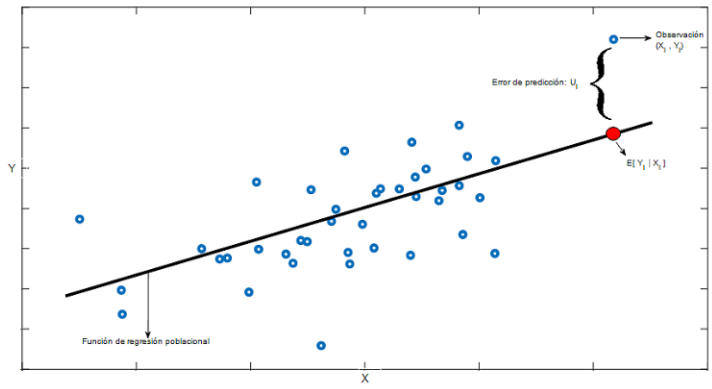

Figura 5. La regresión poblacional. Fuente: Elaboración propia.

Por otro lado, y aquí viene lo bueno, así como tenemos la recta en negro como una posibilidad de $E(Y_i|X_i)$, bien podríamos tener muchas más rectas. En la figura 6 se muestran varias posibilidades en gris. Cada una de ellas nos darían predicciones diferentes para Y_i, es decir los puntos en verde, rosa, rojo, naranja y azul. Y entonces la pregunta es obvia ¿Cuál de todas estas líneas rectas es la mejor? Pues bien, necesitamos una regla que nos permita elegir entre ellas. La forma de elegir entre infinitas posibilidades viene dada por la optimización de lo que en estadística se le conoce como «Función de Pérdida», denotada comúnmente por la letra L.

Por mucho, la función de pérdida más popular en estudios empíricos es la del error cuadrático, también conocida en estadística como función de pérdida del Error Cuadrático Medio. Para la observación i, tendríamos $L(u) = \hat{u}_i^2$.

Es un hecho que no solo nos interesa esa única observación, sino todas las demás también. Entonces tenemos que una forma de cuantificar todos nuestros errores de predicción punto a punto, vendría dado por $L(u) = \sum_{i=1}^{N} \hat{u}_i^2$. Evidentemente nos gustaría equivocarnos lo menos posible, por lo que el problema de optimización a resolver sería $\min_{\Theta} L(u)$, donde Θ es el vector de parámetros que definirá nuestra función de regresión poblacional.

Solo queda matemática por hacer; un sencillo problema de optimización sin restricciones. Suponga que la función de regresión poblacional viene dada por la ecuación de la línea recta $\beta_0 + \beta_1 X_i$. Entonces nuestro modelo de regresión lineal simple es

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{A.1}$$

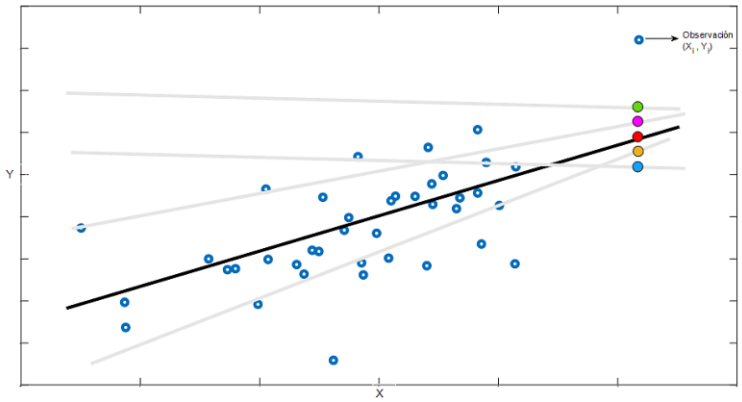

Figura 6. Posibles rectas de regresión. Fuente: Elaboración propia.

donde β_0 y β_1 son los parámetros desconocidos y habremos de estimarlos. En estadística denotamos con el gorrito, $\hat{}$, al estimador del parámetro. No se complique, por ahora entiéndase como un valor supuesto del parámetro. Nuestro objetivo, de una u otra forma, siempre será encontrar estimadores para nuestros parámetros. Teniendo $\hat{\beta}_0$ y $\hat{\beta}_1$, será fácil encontrar nuestro error de estimación aproximado, es decir $\hat{u}_i$. Este recibe el nombre de residual y viene dado por

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i. \tag{A.2}$$

Las siguientes subsecciones ofrecen dos formas de estimar los parámetros.

Mínimos Cuadrados Ordinarios

El método de mínimos cuadrados ordinarios resuelve el problema de optimización que hemos comentado con anterioridad, es decir

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} \hat{u}_i^2 \equiv \min_{\beta_0,\beta_1} \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2. \tag{A.3}$$

La suma de residuales al cuadrado es minimizada por las dos condiciones de primer orden:

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{N} Y_i - N\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{N} X_i = 0, \tag{A.4}$$

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}_1} = \sum_{i=1}^{N} Y_i \, X_i - \hat{\beta}_0 \sum_{i=1}^{N} X_i - \hat{\beta}_1 \sum_{i=1}^{N} X_i^2 = 0. \tag{A.5}$$

Resolviendo para $\hat{\beta}_0$ y para $\hat{\beta}_1$, se obtienen los estimadores de MCO dados por

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2},$$

donde $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

Máxima verosimilitud

Como hemos comentado, existen más formas de estimar el modelo (A.1). La otra forma muy socorrida en estadística y econometría es mediante el método de máxima verosimilitud. Como explicar este método no es ya tan sencillo como con MCO, solo presentamos aquí las ideas principales.

Vamos a suponer que cada uno de nuestros errores, U_i, son independientes e idénticamente distribuidos (*iid*) $\mathcal{N}(0, \sigma^2)$. El lector irá entendiendo que, conforme se requiera en un problema en específico, será relevante estudiar formas de estimación que permitan levantar alguno de los supuestos *iid*. Por el momento, en aras de solo presentar la idea general, nos mantenemos con este supuesto simple.

Huelga decir que este supuesto permite a estadísticos y econometristas derivar distribuciones de estimadores y otros estadígrafos[22] requeridas para hacer inferencia estadística.[23]

En cursos básicos de probabilidad se prueba que una combinación lineal de variables aleatorias normales es en sí misma una variable aleatoria normal. Por lo tanto, es fácil probar que $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$. Podemos entonces escribir la función de densidad conjunta de los errores como:

$$f(u_1, u_2, \ldots, u_N; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right\}.$$

Para obtener la función de verosimilitud, necesitamos hacer la transformación $u_i = Y_i - \beta_0 - \beta_1 X_i$. Note que el jacobiano en esta transformación es 1. Entonces

[22] ¿Qué tal la nueva palabreja? Es una joya, relativamente poco usada en México.

[23] También se puede realizar inferencia estadística con el método de MCO pero se requiere algunos cálculos asintóticos cuyas explicaciones distan mucho de la finalidad de este escrito didáctico.

$$f(Y_1, Y_2, \ldots, Y_N; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} \exp\left\{-\sum_{i=1}^{N} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right\}.$$

Tomando el logaritmo de esta verosimilitud, que denotamos l, obtenemos

$$l(\beta_0, \beta_1, \sigma^2) = -\frac{N}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{N} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}.$$

Ahora bien, maximizando la log-verosimilitud con respecto a los parámetros β_0, β_1 y σ^2, se obtienen los estimadores de máxima verosimilitud (MLE).

Sin embargo, para evitar cálculos innecesarios, note que solo el segundo término en la verosimilitud del logaritmo contiene a β_0, β_1 y ese término (sin el signo negativo) ya ha sido minimizado con respecto a los mismos parámetros anteriormente en (A.4) y (A.5) dándonos los estimadores MCO. Por lo tanto, $\hat{\beta}_{0,MLE} = \hat{\beta}_{0,MCO}$ y $\hat{\beta}_{1,MLE} = \hat{\beta}_{1,MCO}$.

Note que esta igualdad solo se mantiene si la distribución de los errores resulta ser normal. Considere el caso de que la distribución de los errores no sea normal, por ejemplo, si la distribución adecuada es una *t-student*. Ahora imagine que el investigador tiene este conocimiento y decide incorporar dicha distribución de una u otra forma en su modelo. Tal vez no resulte tan evidente pero el investigador deberá estimar su modelo vía MV. En este caso en particular, con el trabajo de este investigador tan experimentado, los estimadores $\hat{\beta}_{0,MV}$ y $\hat{\beta}_{1,MV}$ lograrán ser eminentemente diferentes a sus contrapartes vía MCO y en particular, con propiedades estadísticas muy superiores. Estas propiedades son harina de otro costal, de cursos algo más avanzados en esta hermosa área del conocimiento llamado «modelos lineales».

Bibliografía

[1] P. L. Chebyshev, «Sur les valeurs limites des intégrales», *Journal de Mathématiques Pures et Appliquées*, vol. 12, 1887, 177–184, Original work published in Russian; significant for the development of the method of moments and the Chebyshev inequality.

[2] A. De Moivre, *Approximatio ad summam terminorum binomii (a+ b) n in seriem expansi*, 1733.

[3] F. J. Dyson, «A meeting with enrico fermi», *Nature*, vol. 427, 2004, 297, Dyson recounts the anecdote attributed to John von Neumann about fitting an elephant with four parameters and making it wiggle its trunk with five.

[4] R. A. Fisher, «On the mathematical foundations of theoretical statistics», *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, vol. 222, núm. 594-604, 1922, 309–368.

[5] R. Frisch, «Editor's note», *Econometrica*, vol. 1, núm. 1, 1933, 1–4.
[6] F. Galton, *Hereditary genius: An inquiry into its laws and consequences*, D. Appleton, 1870.
[7] C. F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore carolo friderico gauss*, sumtibus Frid. Perthes et IH Besser, 1809.
[8] J. Geweke, J. Horowitz y M. Pesaran, «Econometrics», en *The New Palgrave Dictionary of Economics*, eds. S. Durlauf y L. Blume, Basingstoke Palgrave Macmillan, 2008.
[9] L. P. Hansen, «Large sample properties of generalized method of moments estimators», *Econometrica: Journal of the econometric society*, 1982, 1029–1054.
[10] A. M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes: avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*, Courcier, 1806.
[11] K. Pearson, «Contributions to the mathematical theory of evolution», *Philosophical Transactions of the Royal Society of London A*, vol. 185, 1894, 71–110, Introduction of the method of moments in statistical estimation.
[12] G. U. Yule, «Why do we sometimes get nonsense-correlations between time-series?– a study in sampling and the nature of time-series», *Journal of the royal statistical society*, vol. 89, núm. 1, 1926, 1–63.