MODELOS LINEALES GENERALIZADOS

ISADORE NABI

I. Conceptos Preliminares ____ *1*

II. MLG y su vínculo genético con los modelos de regresión lineal: la teoría como patrón ____ *4*

II.I. Antecedentes Históricos ____ **4**

II. II. Mínimos Cuadrados Ordinarios y Regresión ____ **5**

II. III. Familias Exponenciales ____ **10**

I.IV. Los Componentes del Modelo Lineal Generalizado ____ **10**

I.IV.I. El modelo lineal clásico como punto de partida ____ 10

I.IV. II. La generalización del modelo lineal clásico ____ 11

I.IV.III. Componentes del MLG ____ 12

I.IV.III. I. El componente estocástico ____ 12

I.IV.III. II. El componente sistemático ____ 13

I.IV.III. III. El enlace entre el componente aleatorio y el componente sistemático: el enlace canónico ____ 13

III. Proceso de Ajuste del Modelo ____ *16*

III.I. Fundamento Estadístico-Matemático Preliminar ____ **16**

III.II. Introducción ____ **20**

III. III. Método de los Mínimos Cuadrados de Reponderación Iterativa (IRLS) ____ **21**

III. III. I. Valores Semilla de la Simulación ____ 21

III. III. II. Funcionamiento del Algoritmo IRLS ____ 22

III.III. II.I. Fundamento matemático ____ 22

III.III. II.II. Funcionamiento mecánico del Algoritmo IRLS ____ 23

III.III. II.III. Análisis del funcionamiento del Algoritmo IRLS ____ 23

III.III. II.IV. Estadísticos Suficientes ____ 24

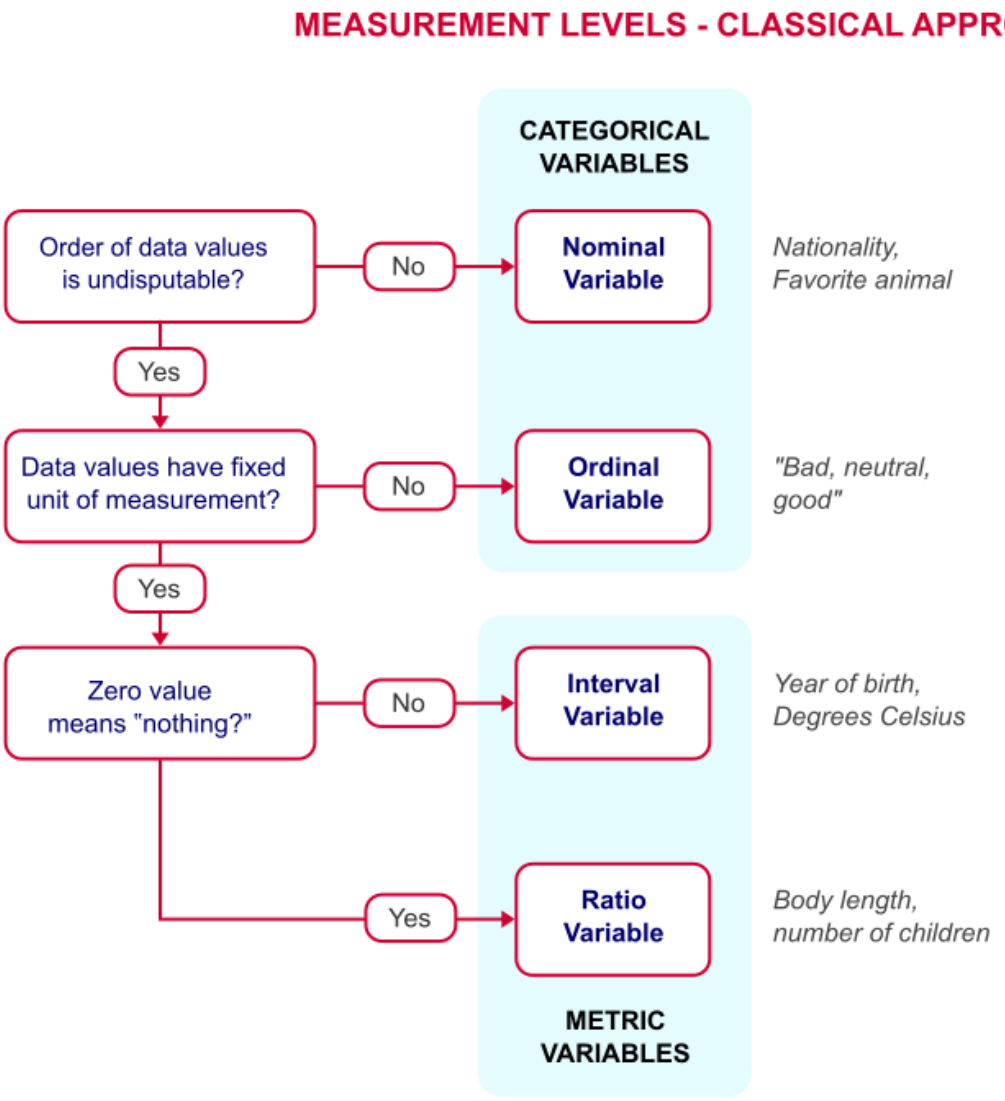
IV. Referencias ____ *25*

I. Conceptos Preliminares

1. *Tipos de variable* según escala de medición (Centro Centroamericano de Población, 2021):
   1.1. *Nominal:* Sus valores sólo se pueden clasificar en clases (o categorías), no se pueden ordenar de pequeño a grande o de menos a más. Ejemplos: sexo, estado civil, profesión, ocupación.
   1.2. *Ordinal:* Sus valores se pueden clasificar en categorías y se pueden ordenar en jerarquías con respecto a la característica que se evalúa. Ejemplos: nivel socioeconómico, Apgar, puntaje Apache de Gravedad cardíaca, clase social, lugar en la clase.

1.3. *De intervalo:* Sus valores tienen un orden natural, es posible cuantificar la diferencia entre dos valores de intervalo. Generalmente tienen unidad de medida. Una variable de intervalo es *discreta* cuando sólo puede tomar un valor entero (por ejemplo: número de hijos, veces que se consultó al establecimiento de salud); o bien es *continua* si puede tomar cualquier valor en un intervalo (por ejemplo.: peso, talla, índice de masa corporal, etc.).

1.4. *De proporción:* El cero representa la ausencia de la característica que se evalúa. Ejemplos: costo por atención, adecuación peso(edad), etc.



Fuente: (van den Berg, 2021).

2. *Nivel de una variable:* Como se señala en (AMERICAN PSYCHOLOGICAL ASSOCIATION, 2021), en el contexto de los diseños experimentales, es la cantidad, magnitud o categoría de la variable independiente (o de un conjunto de ellas) que está siendo estudiada. Por ejemplo., si un investigador está evaluando el efecto del alcohol en la cognición, cada valor específico de alcohol incluido en el estudio es un nivel (*i.e.*, 0.0 oz, 0.5 oz, 1.0 oz, 1.5 oz). Complementariamente, (Online Stat Book, 2021) señala que, si un experimento compara un tratamiento experimental con un tratamiento de control, entonces la variable independiente (tipo de tratamiento) tiene dos niveles: experimental y control. Si en un experimento se comparan cinco tipos de dieta, entonces la variable independiente (tipo de dieta) tiene cinco niveles. En general, el número de niveles de una variable independiente es el número de condiciones en las que la variable independiente se evalúa.
3. *Factor:* Como señala la (AMERICAN PYSCHOLOGICAL ASSOCIATION, 2021), un factor puede tener los siguientes significados:

    3.1. Cualquier cosa que contribuya a un resultado o tenga una relación causal con un fenómeno, evento o acción.
    3.2. Una influencia subyacente que explica en parte las variaciones en el comportamiento individual.
    3.3. En el análisis de varianza y otros procedimientos estadísticos, una variable independiente.
    3.4. En el análisis factorial, una variable latente subyacente no observable que se piensa (junto con otros factores) como responsable de las interrelaciones entre un conjunto de variables.
    3.5. En matemáticas, un número que se divide sin resto en otro número.

Las definiciones de interés en esta sección de la investigación son las definiciones 3.1, 3.2 y 3.3.

4. *Covariable:* Como señala (Allen, 2017, págs. 282-283), una covariable es una variable continua que se espera que cambie ("varíe") con ("co") la variable de salida/resultado/variable dependiente del estudio. En general, una covariable puede referirse a cualquier variable continua que se espera esté correlacionada con la variable de salida de interés.
5. *Variables Métricas:* Como señala (van den Berg, 2021), este es el nombre que reciben las variables que pueden ser de escala de intervalo (que a su vez pueden ser discretas o continuas) o de razón.
6. *Regresión Logística:* Como señala (AMERICAN PYSCHOLOGY ASSOCIATION, 2021), es una forma de análisis de regresión usada cuando la variable independiente (o variable de salida) sólo puede asumir uno de

dos valores categóricos (por ejemplo, aprobar o reprobar) y las predictoras o variables independientes pueden ser tanto categóricas como continuas. Complementariamente, señala (TalkStats, 2011) que la regresión logística multinomial, que es la forma más general de regresión logística, es utilizada para determinar aquellos factores que afectan la presencia o ausencia de una característica cuando la variable dependiente tiene tres o más niveles.

II. MLG y su vínculo genético con los modelos de regresión lineal: la teoría como patrón

II.I. Antecedentes Históricos

Como se señala en (Gujarati & Porter, 2010, pág. 15), fue Francis Galton quien acuñó el término "regresión". En "Family Likeness in Stature", Proceedings of Royal Society, Londres, vol. 40, 1886, pp. 42-72". Ahí planteó que, a pesar de la tendencia de los padres de estatura alta a procrear hijos altos y los padres de estatura baja, hijos bajos, la estatura promedio de los niños de padres de una estatura determinada tendía a desplazarse, o "regresar", a la estatura promedio de la población total. En otras palabras, la estatura de los hijos de padres inusualmente altos o inusualmente bajos tiende a dirigirse a la estatura promedio de la población. Esta ley de regresión universal de Galton fue confirmada por su discípulo y amigo Karl Pearson (junto con A. Lee) en "On the Laws of Inheritance", Biometrika, vol. 2, noviembre de 1903, pp. 357-462. Ahí, se reunen más de mil registros de estaturas de miembros de grupos familiares. Pearson descubrió que la estatura promedio de los hijos de un grupo de padres de estatura alta era menor que la estatura de sus padres, y que la estatura promedio de los hijos de un grupo de padres de estatura baja era mayor que la estatura de sus padres; es decir, se trata de un fenómeno mediante el cual los hijos altos e hijos bajos "regresan" por igual a la estatura promedio de todos los demás. En palabras de Galton, se trata de una "regresión a la mediocridad". La definición moderna de regresión consiste en el "(…) estudio de la dependencia de una variable (variable dependiente) respecto de una o más variables (variables explicativas) con el objetivo de estimar o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las segundas." (Gujarati & Porter, 2010, pág. 15).

Sin embargo, "A pesar de que el análisis de regresión tiene que ver con la dependencia de una variable respecto de otras variables, esto no implica causalidad necesariamente. En palabras de Kendall y Stuart: "Una relación estadística, por más fuerte que y sugerente que sea, nunca podrá establecer una conexión causal nuestras

ideas de causalidad deben provenir de estadísticas externas y, en último término, de una u otra teoría." (…) M. G. Kendall y A. Stuart, The Advanced Theory of Statistics, Charles Griffin Publishers, Nueva York, 1961, vol. 2, cap. 26, p. 279." (Gujarati & Porter, 2010, pág. 19). Profundizando un poco más en ello, (Ritchey, 2002, pág. 522) señala que "La existencia de una correlación tan solo denota que las puntuaciones de las dos variables varían de manera conjunta y sistemática en un patrón predecible. Este descubrimiento por sí mismo no establece causalidad entre las variables. Muchas correlaciones son espurias. Una correlación espuria es aquella que es *conceptualmente falsa, sin sentido o teóricamente sin sentido*, lo cual se ilustra por la correlación entre (…) la tasa de delito en los barrios de la ciudad y la composición racial de una comunidad. Existe una correlación positiva entre el porcentaje de la población minoritaria (por ejemplo, afroamericanos) que viven en barrios y las tasas de crimen. Es decir, para una muestra de comunidades, aquellas con un alto porcentaje de afroamericanos tienden a presentar altas tasas de delito. No obstante, ello sugiere que los afroamericanos son más propensos al comportamiento delictivo, y de hecho, los racistas a menudo citan tal estadística. Esta correlación, sin embargo, resulta espuria. Las tasas de delito son altas en los barrios *pobres* sin tener en cuenta su composición racial, y una parte desproporcionada de los barrios minoritarios son pobres. Es más, la relación entre pobreza y composición racial se debe al racismo, no a la raza biológica. Es decir, ser pobre no tiene nada que ver con la genética. Es la herencia racista de Estados Unidos la que contribuye al hecho de que una parte desproporcionada de los afroamericanos vivan en pobreza, lo cual, a su vez, es un buen predictor de las tasas de delito."[1]

II. II. Mínimos Cuadrados Ordinarios y Regresión

El método de mínimos cuadrados ordinarios y el modelo de regresión lineal no son sinónimos. Como señala (Bhuptani, 2020), hay que comenzar por resaltar primero la diferencia existente entre *regresión lineal* y *ajuste de curvas*. Tener un conjunto de puntos y desear dibujar una curva (línea) a través de ellos que se ajuste lo mejor

[1] A la explicación anterior hay que añadir que no es el racismo por sí mismo el que genera un nexo entre pobreza y composición racial (al menos no entendido como actitud ideológica frente a las personas afro-descendientes), sino que es la exclusión económica y financiera a la que en general se enfrentan los miembros de la sociedad desprovistos de medios de producción, la cual a su vez se agudiza particularmente con los afro-descendientes dadas las condiciones históricas de esclavitud formal, informal y de marginación social en general a la que los distintos imperios que han existido a lo largo de los diversos modos de producción social han sometido a los pueblos africanos desde los tiempos de la antigua Grecia hasta nuestros días. Merece la pena mencionar, en el contexto del movimiento *Black Lives Matters*, que existen dificultades no triviales para delimitar a qué nos referimos con "afro-descendientes", tomando en cuenta que en 1987 los investigadores Rebecca Cann, Stoneking y Wilson demostraron que el Homo sapiens se originó en África entre 140,000 y 290,000 años atrás y migró de allí al resto del mundo, sustituyendo a los humanos arcaicos; véase (Haskett, 2014). Sin embargo, para fines de este análisis tómese de punto de partida la época en que las comunidades primitivas ya estaban bien definidas.

posible a los mismos es un problema puramente geométrico, es decir, los ejes x e y no tienen interpretación, puesto que lo que en otro contexto serían *datos*, en este son meramente puntos en el espacio cartesiano. Por su parte, la regresión lineal es una inferencia estadística sobre un problema concreto de la realidad. Los valores de y se interpretan según el contexto analítico en que se encuentre el investigador[2] y con ello se transforman en *datos sobre la variable de interés* para estudio mediante modelos estadísticos, mientras que los valores de x se transforman en *datos adicionales* que se tiene sobre cada elemento de y que podría ser útil para realizar estimaciones sobre su comportamiento, es decir, transformar los *datos* en *información*, información cuyo carácter debe ser estadísticamente significativo para ser empleado en toma de decisiones relevantes en distintas esferas de la realidad. Cuando se hace una regresión lineal, se está tratando de construir un modelo probabilístico que describa la variable y teniendo en cuenta a la variable x, sin embargo, existen múltiples formas de realizar esto. Un modelo lineal supone que y tiene una media diferente para cada valor posible de x. Así, el conjunto de estos valores medios sigue una línea recta con una cierta intersección y una cierta pendiente. Como con cualquier problema de inferencia estadística, se estiman los parámetros desconocidos utilizando la estimación de máxima verosimilitud. Sin embargo, como en este caso los parámetros desconocidos son una intersección y una pendiente, el resultado final de la estimación de máxima verosimilitud es básicamente que se está eligiendo una línea recta que se ajuste mejor a los datos observados, por lo que es así como convergen la regresión lineal con el ajuste de curvas, *i.e.*, la regresión lineal es el resultado de la estimación de máxima verosimilitud del ajuste del modelo, cuando el conjunto de datos tiene un comportamiento lineal.

Una vez planteado lo anterior, como se señala en la fuente citada, es posible pensar en la regresión lineal como una metodología que utiliza la herramienta del ajuste de curvas (en el caso de la regresión lineal simple, específicamente de una línea recta –o de un hiperplano, si es una regresión lineal múltiple–) mediante un conjunto de puntos que llamamos cualitativamente datos. Sin embargo, existen muchas estrategias posibles para ajustar una línea a través de un conjunto de puntos; por ejemplo, en el contexto de la Ciencia de Datos, existen técnicas para entrenar un modelo lineal que no usa mínimos cuadrados lineales, como se señala en (StackExchange Cross Validated, 2017). A nivel de la Estadística Matemática Clásica existen diferentes metodologías para realizar el ajuste de curvas, entre ellas se encuentran:

[2] Con todas las implicaciones que esto posee.

a) Tomar el punto más a la izquierda y el más a la derecha y dibujar una línea entre ellos.
b) Calcular las pendientes de las líneas que conectan cada par de puntos y calcular la pendiente promedio, dibujando una línea con esta pendiente que pase por el punto en el promedio de los valores de x y el promedio de los valores de y.
c) Se puede encontrar la línea para la cual hay un número igual de puntos sobre la línea y debajo de la línea.
d) Es posible dibujar una línea y luego, para cada uno de los puntos de datos, medir la distancia vertical entre el punto y la línea y sumarlos; la línea ajustada sería aquella donde esta suma de distancias es lo más pequeña posible.
e) También se puede dibujar una línea y luego, para cada uno de los puntos de datos, medir la distancia vertical entre el punto y la línea, elevarlos al cuadrado y sumarlos; la línea ajustada sería aquella donde esta suma de distancias es lo más pequeña posible.

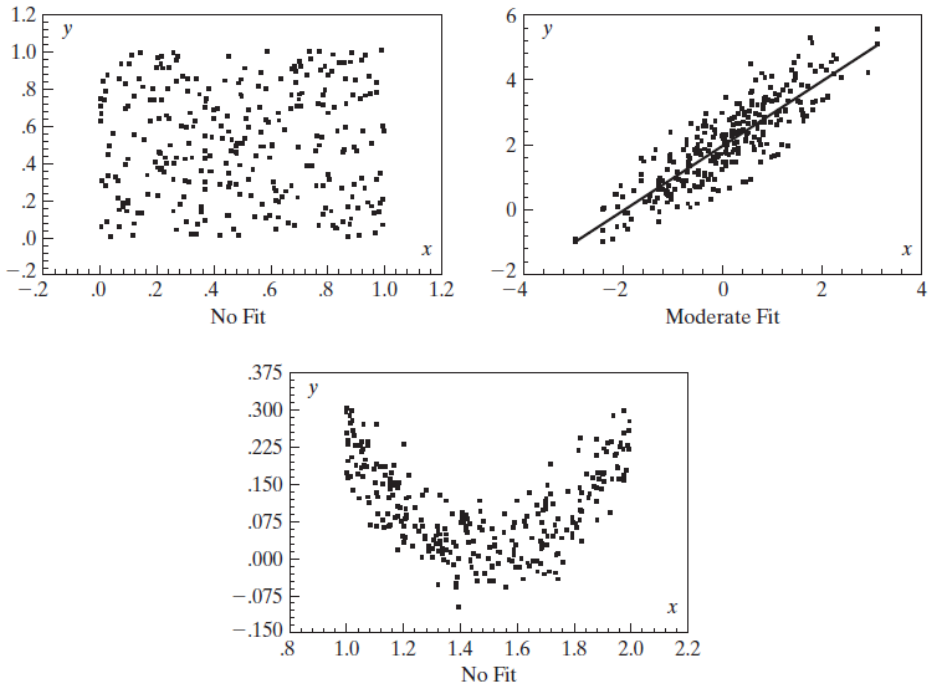
La última estrategia se llama *mínimos cuadrados ordinarios*, de ahora en adelante *MCO*, y su nombre proviene del hecho que se está buscando minimizar la suma de los errores de predicción al cuadrado. Sin embargo, a pesar de que los *MCO* son la técnica más popular que emplea la metodología del Análisis de Regresión, en lo que respecta al ajuste de una línea a través de un conjunto de puntos, cualquiera de las otras estrategias es igualmente válida. Las tres primeras estrategias las inventó el autor a manera de ejemplo y probablemente no funcionen adecuadamente; sin embargo, la cuarta es una estrategia real llamada *desviaciones menos absolutas* y es preferida por algunas personas por sobre mínimos cuadrados.

Cabe preguntarse por qué si no es la única técnica entonces es la más utilizada. La razón es que, al resolver el problema de regresión lineal estadística, una suposición de modelado muy común es que por cada valor posible de x, la cantidad y se distribuye normalmente con una media que es lineal en x. Por lo tanto, la función de verosimilitud es esencialmente un producto de funciones de densidad de probabilidad de la distribución normal. Así mismo, el autor señala que se estiman los parámetros desconocidos (y, por lo tanto, se encuentra la recta de mejor ajuste al conjunto de observaciones) maximizando la función de verosimilitud. Si se observa cómo es el producto de funciones de densidad de probabilidad normales, el lector notará que maximizar esta expresión es equivalente a minimizar la suma de los errores al cuadrado. Es decir, la línea que se obtiene realizando el ajuste de la curva a través de mínimos cuadrados es equivalente a la línea que obtiene realizando una regresión lineal utilizando un modelo distribuido normalmente.

De esta forma, puede observarse que el análisis de regresión es una metodología, mientras que los mínimos cuadrados con una técnica empleada por esta metodología. Mucho menos debe identificarse una regresión lineal con la técnica de mínimos cuadrados, puesto que, por ejemplo, existen distintos tipos de análisis de regresión, entre ellos el análisis de regresión lineal. Sin embargo, los mínimos cuadrados son una de las técnicas posibles en la regresión lineal para encontrar la línea recta de mejor ajuste al conjunto de datos del que se dispone. Así, se presentan a continuación dos figuras, la primera permite ver los diferentes ajustes que un conjunto de datos muestrales puede tener respecto a una recta y cómo esta se convierte en la recta de mejor ajuste, mientras que la segunda permite visualizar la descomposición de cada y_i en la regresión lineal.

The original fitting criterion, the sum of squared residuals, suggests a measure of the fit of the regression line to the data. However, as can easily be verified, the sum of squared residuals can be scaled arbitrarily just by multiplying all the values of y by the desired scale factor. Since the fitted values of the regression are based on the values of **x**, we might ask instead whether *variation* in **x** is a good predictor of *variation* in y. Figure 3.3 shows three possible cases for a simple linear regression model. The measure of fit described here embodies both the fitting criterion and the covariation of y and **x**.

FIGURE 3.3 Sample Data.



Fuente: (Greene, 2012, pág. 79).

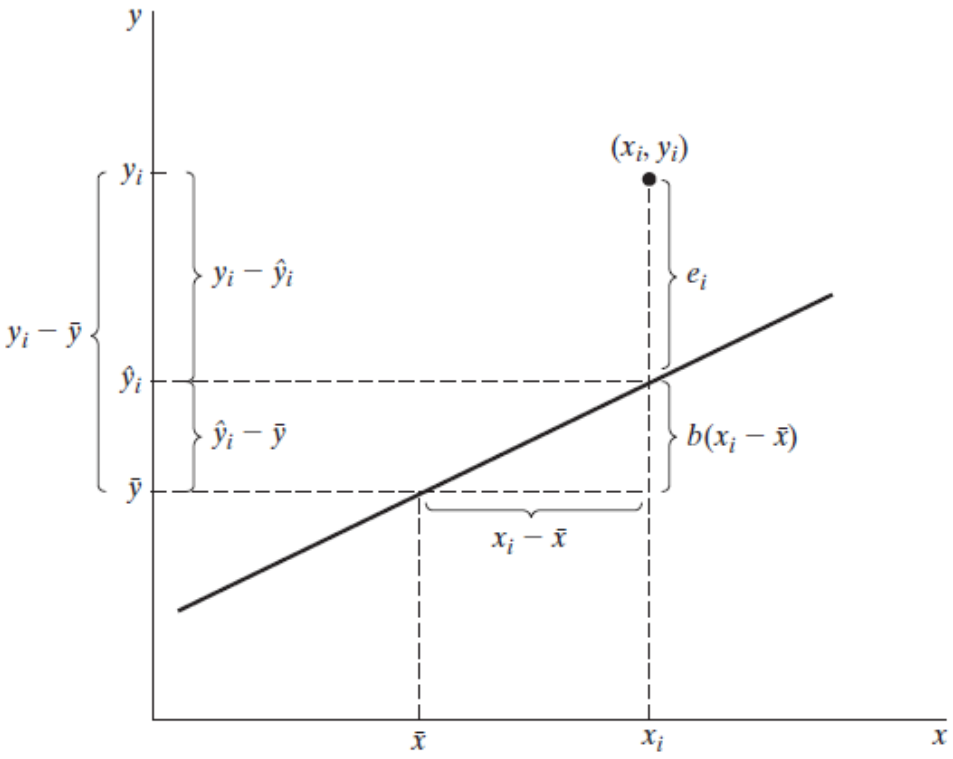


FIGURE 3.4 Decomposition of y_i.

Fuente: (Greene, 2012, pág. 80).

En suma, pueden entenderse los mínimos cuadrados como una técnica en el análisis de regresión para aproximar la solución de sistemas sobre determinados (i.e., conjuntos de ecuaciones en las que hay más ecuaciones que incógnitas) minimizando la suma de los cuadrados de los residuos hechos en los resultados de cada ecuación. Por su parte, los mínimos cuadrados lineales es una técnica de aproximación de funciones lineales a conjuntos de datos por la técnica de mínimos cuadrados. Es un conjunto de formulaciones para resolver problemas estadísticos relacionados con la regresión lineal, incluidas las variantes para residuos ordinarios (no ponderados), ponderados y generalizados (correlacionados). Los métodos numéricos para mínimos cuadrados lineales incluyen la inversión de la matriz de las ecuaciones normales y los métodos de descomposición ortogonal. Por su parte, los mínimos cuadrados ordinarios son una de las 3 técnicas más comunes (más no las únicas) dentro de la familia de técnicas conocida como "Mínimos Cuadrados Lineales", que a su vez pertenece a la familia de técnicas conocida como "Mínimos Cuadrados". Los mínimos cuadrados ordinarios son utilizados en el contexto del modelo clásico de regresión lineal para estimar sus parámetros desconocidos.

En línea con (McCullagh & Nelder, 1989, pág. 4), deben considerarse a las diferentes teorías estadísticas como descripciones de determinados patrones que es

posible identificar que siguen los números en la vida real, patrones los cuales en alguna medida pueden sustituir al conjunto de datos en sí mismos (puesto que estos patrones numéricos describen patrones geométricos, es decir, relativo a las formas que adoptan los fenómenos naturales y/o sociales) a través de determinados valores numéricos concretos en los que se cristalizan dichos patrones. Es por ello que según las características empíricas del conjunto de datos en concreto que se estudie (obtenido de la medición de fenómenos naturales y/o sociales) los parámetros $\boldsymbol{\beta}$ generados por tal conjunto de datos tomarán diferentes valores y precisamente de este hecho empírico es que se formulan las teorizaciones estadísticas-matemáticas que actualmente se conocen como *familias de distribuciones de probabilidad*.

El modelo básico de regresión lineal $y = \alpha + \beta x$ conecta dos variables x e y vía el par de parámetros (α, β) y define una relación entre ambas que describe geométricamente una línea recta.

II. III. Familias Exponenciales

Formalmente, según (Patil & Shorrock, 1965, pág. 94), una familia exponencial se define como aquella familia $\{X_\omega : \omega \in \Omega\}$, en donde Ω es un espacio de variables estocásticas real-evaluadas X_ω en el que cada una de estas tiene asociada una función de densidad (o de masa) de probabilidad F_ω y en el que la derivada de la función respecto a x tiene la forma $dF_\omega(x) = \left\{\frac{e^{\omega x}}{f(\omega)}\right\} dv(x)$ y $\int e^{\omega x} dv(x)$.

En la definición anterior, v es una función que sirve como medida[3] en las integrales de Lebesgue–Stieltjes[4] y es una función no decreciente de variable real, mientras que ω es el conjunto de parámetros que en la notación usual suelen encontrarse como θ.

I.IV. Los Componentes del Modelo Lineal Generalizado

I.IV.I. El modelo lineal clásico como punto de partida

Como señalan (McCullagh & Nelder, 1989, pág. 26), los modelos lineales generalizados son una extensión de los modelos clásicos, por lo que estos últimos representan el punto de partida de la exposición.

Un vector de observaciones y que posee n componentes se asume que es una realización de la variable aleatoria Y cuyos componentes están independientemente distribuidos con medias μ. El componente sistemático del modelo es una especificación del vector μ en términos de un pequeño número de

[3] Como señala (Kolmogórov & Fomin, 1978)

[4] Es una de las generalizaciones posibles de la integral de Riemann y de Stieltjes, fundamentada bajo el marco formal de la Teoría de la Medida fundada por Henri Lebesgue, en sentido en que lo están las integrales de Lebesgue.

parámetros desconocidos $\beta_1, \beta_2, \dots, \beta_p$. En el caso de los modelos lineales ordinarios, esta especificación toma la siguiente forma:

$$\mu = \sum_{1}^{p} x_j \beta_j$$

En la expresión anterior, los β son parámetros cuyos valores son usualmente desconocidos y deben ser estimados a partir del conjunto de datos. Si se indizan las observaciones mediante la letra i, es posible entonces expresar al componente sistemático del modelo de la siguiente manera:

$$E(Y_i) = \mu_i = \sum_{1}^{p} x_{ij} \beta_j \, ; i = 1,2, \dots, n,$$

En la expresión anterior, x_{ij} es el valor de la $j - ésima\ covarianza$ de la observación i, que en este caso es una variable en sí misma (por ello es que se habla de $j - ésima$ covarianza). En notación matricial, en donde μ es una matriz de orden $n \times 1$, X una matriz de orden $n \times p$ y β una matriz de orden $p \times 1$, es posible escribir lo anterior como $\mu = X\beta$, en donde X es la matriz modelo y β el vector de parámetros; la estructura del componente sistemático asume que las covarianzas que perturban a la media son conocidas y pueden ser medidas con efectividad y libre de errores, lo cual también debe verificarse con el conjunto de datos del que se disponga en la medida en que sea posible. Para el caso de la parte estocástica, se asume independencia entre sus elementos y varianza constante de los errores. Estos supuestos son fuertes a nivel teórico y deben verificarse, en tanto sea posible, de los datos mismos.

Algunas especializaciones del modelo lineal clásico asumen supuestos más fuertes (restrictivos) como que los errores de estimación siguen una distribución normal con varianza constante σ^2. Sintetizando lo visto sobre el modelo lineal clásico, este puede ser expresado como $E(Y) = \mu$, donde $\mu = X\beta$.

I.IV. II. La generalización del modelo lineal clásico

Como señalan (Nelder & Wedderburn, 1972, pág. 370), "Los modelos lineales habitualmente incorporan componentes tanto sistemáticos como aleatorios (error), y los errores generalmente se asume que tienen distribuciones normales. La técnica analítica asociada es la teoría de mínimos cuadrados, que en su forma clásica asumía solo un componente de error; las extensiones para errores múltiples se han desarrollado principalmente para el análisis de experimentos diseñados y datos de encuestas. Las técnicas desarrolladas para datos no normales incluyen análisis Probit, donde una variable binomial tiene un parámetro relacionado con una distribución de tolerancia subyacente asumida, y tablas de contingencia, donde la

distribución es multinomial y la parte sistemática del modelo, generalmente multiplicativa."

Así, la generalización del modelo lineal clásico es posible puesto que "En ambos ejemplos hay un aspecto lineal del modelo; por lo tanto, en el análisis Probit, el parámetro p es una función de la tolerancia Y, que en sí misma es lineal sobre la dosis (o alguna función de la misma), y en una tabla de contingencia con un modelo multiplicativo, el logaritmo de la probabilidad esperada se asume lineal al clasificar los factores que definen la mesa. Por tanto, para ambos, la parte sistemática del modelo tiene una base lineal. En otra extensión (Nelder, 1968) se usa cierta transformación para producir errores normales, y se usa una transformación diferente de los valores esperados para producir linealidad. Hasta ahora hemos mencionado modelos asociados con las distribuciones normal, binomial y multinomial (esta última puede considerarse como un conjunto de distribuciones de Poisson con restricciones). Una clase adicional se basa en la distribución χ^2 o gamma y surge en la estimación de los componentes de la varianza a partir de formas cuadráticas independientes derivadas de las observaciones originales. Nuevamente, el componente sistemático del modelo tiene una estructura lineal (…) En esta investigación, nosotros desarrollamos una clase de modelos lineales generalizados, los cuales incluyen todos los ejemplos anteriores, y proporcionamos un proceso unificado para ajustarlos con base en la verosimilitud. Este procedimiento es una generalización del procedimiento bien-conocido descrito por Finney (1952) para máxima verosimilitud en el contexto del análisis Probit" (Nelder & Wedderburn, 1972, págs. 370-372).

Para simplificar la transición del modelo lineal clásico al generalizado, se debe reestructurar sutilmente $E(Y) = X\beta$ para producir la siguiente especificación de tres partes.

I.IV.III. Componentes del MLG

I.IV.III. I. El componente estocástico

Así, con base en (Nelder & Wedderburn, 1972, pág. 371), supóngase que las observaciones que conforman el conjunto de datos pueden ser descrita por una función de densidad (o de masa) de probabilidad π de la siguiente forma:

$$\pi(z;\ \theta, \phi) = \exp\left[\alpha(\phi)\{z\theta - g(\theta) + h(z)\} + \beta(\phi, z)\right]$$

En donde α, g, h y β son conocidas, así como también $\alpha(\phi) > 0$ tal que para un valor fijo de ϕ se tiene una familia exponencial descrita por $\pi(z;\ \theta, \bar{\phi})$. θ representa los parámetros de la distribución de la variable dependiente descrita por el conjunto de observaciones $z \in Z$ y ϕ es un parámetro de dispersión, usualmente asociado a la varianza de las distribuciones de probabilidad, aunque también

puede ser, por ejemplo, el parámetro p para el caso de una distribución gamma. La media de z se denota como μ. Para el caso de vectores de variables aleatorias o vectores estocásticos Y, los componentes de Y tienen distribuciones independientes e idénticas (*iid*) con esperanza matemática $E(Y) = X\beta = \mu$ y varianza constante σ^2 o cualquier otro parámetro de dispersión.

I.IV.III. II. El componente sistemático

Como se señala en (McCullagh & Nelder, 1989, pág. 26), las covarianzas $x_1, x_2, \dots, x_p$ producen un predictor lineal η definido como $\sum_1^p x_j\beta_j$. En este sentido, se señala en (Nelder & Wedderburn, 1972, pág. 372) que las variables independientes pueden ser cuantitativas y producir un único valor para la variable x en el modelo, pueden ser cualitativas y producir un conjunto de valores de x conformado por la opción 0 y la opción 1, o puede ser una mezcla de ambos tipos.

I.IV.III. III. El enlace entre el componente aleatorio y el componente sistemático: el enlace canónico

Con este enlace se garantiza que $\mu = \eta$. Esta generalización introduce, como puede verificarse, un nuevo símbolo η para el predictor lineal y el tercer componente, para luego especificar que μ y η son de hecho idénticos. Si se escribe $\eta_i = g(\mu_i)$, entonces $g(\cdot)$ será llamada *función enlace*.

Como se adelantó, preliminarmente los modelos lineales clásicos estaban restringidos al componente estocástico y a la identidad entre el predictor lineal y la media. Los modelos lineales generalizados permiten dos extensiones:

1) La distribución de probabilidad que sigue el componente estocástico ya no está restringida únicamente a ser normal, sino que puede ser generada por los miembros de una familia de funciones exponenciales diferente de la normal.
2) La función de enlace especificada en el tercer componente puede ser cualquier función monotónica diferenciable.

Complementariamente, acorde con (McCullagh & Nelder, 1989, pág. 31), se debe mencionar que la función enlace relaciona al predictor lineal η con el valor esperado μ de un conjunto de datos y. En los modelos lineales clásicos la media y el predictor lineal eran idénticos, y el enlace identidad es plausible en el sentido de que tanto η como μ pueden tomar valores que pertenecen a los números reales. Sin embargo, cuando se está trabajando con conteos y la distribución de Poisson, se debe tener $\mu > 0$, por lo que el enlace identidad es menos atractivo, en parte porque η puede ser negativa mientras que μ no debe serlo. Modelos de conteo basados en la independencia en datos de clasificación cruzada conduce naturalmente a efectos multiplicativos, y esto está expresado en el *log-enlace* de la

forma $\eta = \log(\mu)$, cuya inversa es $\mu = e^{n}$. Además, los efectos aditivos que contribuyen a η se convierten en efectos multiplicativos que contribuyen a μ, y μ es necesariamente positiva. El tratamiento matemático dado a cinco distribuciones de probabilidad, en el contexto de los modelos lineales generalizados, se presentan a continuación.

Table 2.1 *Characteristics of some common univariate distributions in the exponential family*[†]

	Normal	*Poisson*	*Binomial*	*Gamma*	*Inverse Gaussian*
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m, \pi)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
Range of y	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0, \infty)$	$(0, \infty)$
Dispersion parameter: ϕ	$\phi = \sigma^2$	1	$1/m$	$\phi = \nu^{-1}$	$\phi = \sigma^2$
Cumulant function: $b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y; \phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$-\log y!$	$\log\binom{m}{my}$	$\nu\log(\nu y) - \log y$ $-\log\Gamma(\nu)$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$
$\mu(\theta) = E(Y; \theta)$	θ	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
Canonical link: $\theta(\mu)$	identity	log	logit	reciprocal	$1/\mu^2$
Variance function: $V(\mu)$	1	μ	$\mu(1 - \mu)$	μ^2	μ^3

[†]The mean-value parameter is denoted by μ, or by π for the binomial distribution.
The parameterization of the gamma distribution is such that its variance is μ^2/ν.
The canonical parameter, denoted by θ, is defined by (2.4). The relationship between μ and θ is given in lines 6 and 7 of the Table.

Fuente: (McCullagh & Nelder, 1989, pág. 30).

III. Proceso de Ajuste del Modelo

III.I. Fundamento Estadístico-Matemático Preliminar

El ajuste de una simple relación lineal entre los elementos x y los elementos y requiere que sea utilizado un par de valores en particular (a, b), seleccionado del conjunto de todos los posibles pares de valores que pueden adoptar los parámetros (α, β), que genere valores estimados $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ que se ajusten (en el sentido de generar una figura geométrica -en este caso una línea recta-) mejor a los valores observados $y_1, y_2, \dots, y_n$.

Con la finalidad de cuantificar las diferencias en el ajuste entre los valores estimados $\hat{y}_n$ y los valores observados y_n, se deben medir las distancias entre tales conjuntos de valores. En general, estas distancias (referidas en textos como el de McCullagh y Nelder como discrepancias) buscan cuantificarse con la finalidad de optimizar el proceso de selección del patrón geométrico teórico que mejor describa el patrón observado descrito por el conjunto de datos. Para cuantificar tales distancias se pueden utilizar diferentes tipos de métricas (criterios de medición), selección que a su vez responderá a las características específicas del conjunto de datos, sobre lo cual se abordarán brevemente a continuación algunos aspectos.

Cada uno de los tipos de métrica que es posible utilizar para medir distancias se conoce, en el contexto del Análisis Matemático en general y del Análisis Real en particular (campo al que pertenece a su vez la Estadística Matemática), como *función distancia* en espacios topológicos equipados con métrica. Los espacios topológicos son estructuras matemáticas en las cuales existen reglas de agrupación o *topología* de los elementos de un determinado conjunto en subconjuntos, en donde tales reglas de agrupación expresan las relaciones entre los elementos que conforman el conjunto y a partir de las cuales se generan los demás tipos de relaciones entre tales elementos, las cuales geométricamente representan sus distancias relativas, como se verifica en los anexos correspondientes al estudio de los grafos desde la perspectiva de los isomorfismos. Por otro lado, la métrica o función distancia es la función utilizada para medir las distancias absolutas entre los elementos de un conjunto o, para este caso, las distancias absolutas de las observaciones que conforman un conjunto de datos. Los distintos tipos de métrica son siempre generalizaciones de la métrica de espacios euclidianos, expresada en su forma básica en el ampliamente conocido *teorema de Pitágoras*.

Con la finalidad de ofrecer una exposición integral de lo presentado por (McCullagh & Nelder, 1989, pág. 5) en relación a las normas L_1 y L_2, específicamente para comprender cómo es que estos conceptos matemáticos son el fundamento del ajuste del modelo de regresión utilizado en cuanto son fundamento teórico y aplicado de la métrica empleada en las mediciones de

cualquier índole, se abordarán sintéticamente algunos conceptos del Análisis Matemático y el Álgebra Abstracta.

El primer concepto a estudiar es el de *función convexa*. Como señalan (Kolmogórov & Fomin, 1978, pág. 140), una *funcional*[5] *convexa* no negativa ρ, definida sobre un espacio lineal real L, se llama convexa si cumple las siguientes condiciones:

1) $\rho(x+y) =\leq \rho(x)+\rho(y)$, para todo x, y que pertenece a L.
2) $\rho(\alpha x) = \alpha\rho(x)$, para todo $\alpha \geq 0$
3) No se admite que el valor de $\rho(x)$ es finito para todo x que pertenece a L, es decir, se admite el caso en que $\rho(x) = +\infty$ para algunos x que pertenecen a L.

El segundo concepto a estudiar es el de *norma*. Como señala (Kolmogórov & Fomin, 1978, pág. 149), una funcional convexa finita ρ, definida sobre L, se llama *norma* cuando verifica las siguientes dos condiciones, adicionales a las de convexidad:

1) $\rho(x) = 0$, sólo si $x = 0$.
2) $\rho(\alpha x) = |\alpha|\rho(x)$, para todo α.

Sintetizando las condiciones vistas en el primer y segundo concepto, se obtienen:

1) $\rho(x) \geq 0$, con la particularidad de que $\rho(x) = 0$ sólo si $x = 0$.
2) $\rho(x+y) \leq \rho(x)+\rho(y)$, para todo x, y que pertenece a L.
3) $\rho(\alpha x) = |\alpha|\rho(x)$, para todo α.

Generalizando las tres condiciones anteriores, se tiene

$$\left\|\sum_{k=1}^{n} \alpha_k x_k\right\| \leq \sum_{k=1}^{n} |\alpha_k|\|x_k\|$$

Así, en el contexto de los *espacios topológicos lineales* (conocidos en cursos de matemática elemental como espacios vectoriales) es posible, en conjunto con el cumplimiento de otras condiciones de carácter más general (global) que no se especificarán aquí, derivar de la función norma una función métrica. La importancia de que la métrica sea inducida por una norma, a nivel de los espacios euclidianos y sus generalizaciones más usadas (los espacios de Hilbert), proviene del hecho de que así se garantiza que la función métrica posea la característica de ser invariante ante traslaciones, *i.e.*, una *métrica invariante ante traslaciones* a la que suele hacer alusión la literatura bajo el nombre de *principio de traslación invariante*, así como también la tercera propiedad relativa a los escalares α. Lo que en

[5] Un funcional es una generalización del concepto de función, específicamente es una función de funciones.

términos cotidianos significa que una función métrica sea invariante ante traslaciones es que dicha función, al realizar mediciones de cualquier índole sobre algún objeto localizado en la estructura matemática en cuestión, no arrojará mediciones diferentes cuando el objeto sea trasladado de un lugar a otro dentro de la misma estructura matemática, que para este caso son los espacios topológicos lineales antes mencionados.

Como puede verificarse en señala (Kolmogórov & Fomin, 1978, pág. 52)[6], las características de la función métrica d de espacios euclidianos, conocida también como *función distancia* d de espacios euclidianos, adicionales son las siguientes:

1) Tomando para los elementos de un conjunto arbitrario
$$d(x,y) = \begin{cases} 0, si\ x = y \\ 1, si\ x \neq y \end{cases}$$
se obtendrá, evidentemente, un espacio métrico que puede ser denominado espacio de puntos aislados.
2) El conjunto de los números reales con la distancia $d(x,y) = |x - y|$ forma el espacio métrico R^1.
3) El conjunto de grupos ordenados de n números reales $x = (x_1, x_2, \dots, x_n)$ con la distancia $d(x,y) = \sqrt{\sum_{k=1}^{n}(y_k - x_k)^2}$ se denomina *espacio aritmético euclídeo de n dimensiones* R^n. Esta condición es conocida también como cumplimiento del *axioma triangular*.

Como se verifica en (Lipschutz, 1992, pág. 51)[7], las tres condiciones anteriores pueden expresarse de forma general como se presenta a continuación:

a) $d(x,y) = 0$, (para $x = y$)[8]

b) $d(x,y) = \rho(y,x) > 0$, (para $x \neq y$)

c) $d(x,y) \leq \rho(x,z) + \rho(z,y)$, (para $z \neq x, y$)[9]

d) $d(x,y) \geq 0$, (positividad, deducida de las tres propiedades anteriores).

---

[6] La obra citada no hace diferenciación entre la función métrica y la función norma en términos de notación, puesto que para ambas utiliza ρ. En esta investigación hemos sustituido la ρ métrica por d, ello con la finalidad de mostrar la mecánica operativa de cómo la métrica es inducida por una norma.

[7] Aunque su formato de presentación se ha generalizado por cuenta propia para estandarizar notación y nivel de abstracción matemática con las referencias tomadas de (Kolmogórov & Fomin, 1978) (que es el libro de referencia que se usa en esta investigación en lo relativo al Análisis Matemático). Ello con la finalidad de ofrecer mayor claridad y fluidez en la narrativa.

[8] También conocida como *Identidad de los Indiscernibles*.

[9] Esta condición es precisamente la que garantiza que la métrica o función distancia sea invariante ante traslaciones.

El concepto de espacio real R^n tiene su generalización natural en el concepto de espacios L_p. Como se verifica en (Wikipedia, 2021), los espacios L_p son espacios de funciones (*i.e.*, espacios más abstractos a los usuales en donde las coordenadas están dadas en términos de funcionales y no de números reales) definidos como generalización natural de la norma ρ para espacios vectoriales de dimensión finita. Estos espacios también son conocidos como *espacios de Lebesgue*. De ahí que se hable de norma L_1, norma L_2 y hasta norma L_p. ¿Cómo exactamente es que la métrica es inducida por una norma?

Sean $x, y, z \in V$, donde x, y, z son variables, V un espacio topológico lineal, $d(x, z)$ una métrica entre un punto x y un punto z, y sea también $\rho(x - y)$ la función norma de interés. Mostrar cómo una métrica es inducida por una norma es equivalente a demostrar que al igualar matemáticamente la expresión $d(x, z)$ con la expresión $\rho(x - y)$ y desarrollar la expresión resultante, es posible obtener un resultado, conocido también como *métrica inducida* (por una norma), que expresa una función métrica que posee las propiedades de una función métrica usual, lo cual se verifica si partiendo de la norma de interés es posible arribar a métrica invariante ante traslaciones[10]. Esto se muestra a continuación.

$$
\begin{aligned}
d(x, z) = \rho(x - y) &= \rho\big(x + (-z)\big) \\
&= \rho\big((x - y) + (y - z)\big) \\
&\leq \rho(x - y) + \rho(y - z) \\
&= \rho(x - y) + |-1|\rho(y - z) \\
&= \rho(x, y) + \rho(y, z) \\
&= d(x, y) + d(y, z)
\end{aligned}
$$

Así, como señalan (McCullagh & Nelder, 1989, pág. 5), seleccionar el modelo estadístico óptimo para determinado conjunto de datos se hace bajo el criterio de cuál es el que genera un conjunto de datos estimado más cercano al conjunto de datos observado (los que han sido capturados a través de mediciones a fenómenos de la realidad), lo cual equivale a determinar la discrepancia cuantitativa (en términos de sus distancias dentro del espacio de Lebesgue en el que se estudian) que existe entre cada uno de los elementos del conjunto de datos estimado y los del conjunto de datos observado, lo cual los econometristas suelen conocer a nivel empírico como *residuo de la regresión* y a nivel teórico como *término de perturbación estocástica* o *término de error*. Esta discrepancia se cuantifica mediante alguna función norma, cuya selección dependerá de las características específicas del

[10] Véase (Perry, 2014).

conjunto de datos que se estudie (cuyas características vienen conferidas por las del fenómeno que cuantifican y por el diseño mismo del instrumento de medición con el que se capturaron estos datos); sin embargo, estas normas poseen todas limitantes en común que condiciona su uso, las cuales se expondrán más adelante.

III.II. Introducción

Como señalan (McCullagh & Nelder, 1989, pág. 5), existen desde la norma L_1 definida como $S_1(y, \hat{y}) = \sum|y - \hat{y}|$ hasta la norma L_p definida como $S_\infty(y, \hat{y}) = \max_i|y - \hat{y}|$. Sin embargo, los espacios que usualmente son de más interés aplicado, así como también aquellos en que se realiza el ajuste clásico por mínimos cuadrados, son los espacios L_2 normados, conocidos también a nivel estadístico como espacios de desviaciones cuadráticas, que son los espacios aritméticos euclidianos $n - dimensionales$ antes definidos a nivel métrico. A continuación, se define la norma de estos espacios.

$$S_2(y, \hat{y}) = \sum (y_i - \hat{y}_i)^2$$

Como se mencionó, la validez de las estimaciones realizadas tiene determinados elementos en común con independencia del orden de la norma que se utilice, orden que indica precisamente la potencia a la que se eleva la diferencia entre las localizaciones de los elementos de los conjuntos de datos (del estimado o generado con la regresión y del observado). Estos elementos en común son tres:

1) En primer lugar, todas las mediciones (expresadas en las observaciones) han sido realizadas bajo la misma escala física.
2) Las observaciones son independientes entre sí o, al menos, "(…) ellas son en algún sentido intercambiables, lo que un trato imparcial de los componentes." (McCullagh & Nelder, 1989, pág. 5). Esta noción puede generalizarse a lo que es posible concebir como *intercambiabilidad estocástica*, que generaliza la noción de *independencia estocástica*.
3) Cada una de las desviaciones debe ser independiente del valor esperado del conjunto de observaciones.

Como se deduce de lo anterior, cada una de las normas $L\ p - ésimas$ se corresponde con un determinado criterio estadístico, específicamente con la potencia a la que se elevan las distancias entre los valores observados y los valores estimados, puesto que así se realiza la medición de los residuos o errores en el contexto de los modelos de mínimos cuadrados. En este sentido, (Wikipedia, 2021) señala que el método IRLS se utiliza para encontrar las estimaciones de máxima verosimilitud de un modelo lineal generalizado como una forma de mitigar la influencia de valores atípicos en un conjunto de datos normalmente distribuido.

Por ejemplo, minimizando los errores mínimos absolutos (expresado en la norma L_1) en lugar de los errores de mínimos cuadrados (expresado en la norma L_2).

III. III. Método de los Mínimos Cuadrados de Reponderación Iterativa (IRLS)

Como se adelantó, el proceso de ajuste del modelo se lleva a cabo mediante el método de máxima verosimilitud. En este sentido, se señala en (Nelder & Wedderburn, 1972, págs. 372-373) que la solución a las ecuaciones de máxima verosimilitud generadas para el caso de $n-$ é$simas$ variables estocásticas contenidas dentro del vector estocástico Y es equivalente a un procedimiento iterativo por mínimos cuadrados ponderados con una función de ponderación:

$$w = \frac{\left(\frac{d\mu}{dY}\right)^2}{V}$$

En la expresión anterior, w son las ponderaciones, $\frac{d\mu}{dY}$ es el diferencial del valor esperada respecto de Y, mientras que V es la varianza de las observaciones, en donde μ, Y y V expresan estimaciones derivadas del conjunto de datos con el que se cuenta. Además, Y (que es el conjunto de datos pronosticado, no el observado) es expresada en el método IRLS a través de la expresión $y = Y + \frac{z-\mu}{\frac{d\mu}{dY}}$, conocida como *working Probit*.

III. III. I. Valores Semilla de la Simulación

Como señalan (Nelder & Wedderburn, 1972, pág. 374)En la práctica, podemos obtener un buen procedimiento de partida para la iteración de la siguiente manera: tomar como primera aproximación $\mu = z$ y calcular Y a partir de ese valor; luego calcúlese w como se definió y establézcase $y = Y$. Luego obténgase la primera aproximación a los β's por regresión. El método puede necesitar una ligera modificación para tratar con valores extremos de z. Por ejemplo, con la distribución binomial probablemente será adecuado reemplazar instancias de $z = 0$ o $z = n$ con $z = \frac{1}{2}$ y $z = -\frac{1}{2}$ en aquellos casos en los que, por ejemplo, las transformaciones Probit y Logit, $\mu = 0$ o $\mu = n$ conducirían a obtener valores infinitos para Y.

III. III. II. Funcionamiento del Algoritmo IRLS

III.III. II.I. Fundamento matemático

Como se señala en (Burrus, 2021), La distancia de una observación cualquiera respecto a la medida de tendencia central de su distribución se conoce como desviación. En el contexto de los modelos predictivos, las desviaciones se conocen como residuos, denotados mediante la letra e. Estos residuos en el contexto de la regresión lineal se estiman matricialmente de la forma:

$$e = Ax - b \quad (1)$$

Como ya es sabido, la ecuación de la norma puede tomar la siguiente forma:

$$\|e\|_p = \left(\sum_n |e(n)|^p\right)^{1/p} \quad (2)$$

Una regresión consiste, en general y sintéticamente, en ajustar un conjunto de datos a una función, la función de mejor ajuste, mediante la minimización de los residuos elevados a alguna potencia, usualmente al cuadrado. Si existe una solución óptima o aproximadamente óptima para minimizar el error de la primera ecuación matricial, entonces es posible alcanzar este objetivo mediante la minimización de la segunda ecuación, correspondiente a la ecuación de la norma del residuo.

Se sabe que la fórmula para encontrar el residuo cuadrado mínimo ponderado es:

$$\|W_e\|_{p=2}^2 = \sum_n w_n^2 |e(n)|^2 \quad (3)$$

Así, existen demostraciones matemáticas que prueban que minimizar la norma del residuo equivale a minimizar la siguiente ecuación:

$$\|e\|_p = \sum_n (w(n)^2 |e(n)|^2)^{1/2} \quad (4)$$

Para realizar exitosamente la minimización de las distancias a través de la norma, se necesita encontrar un valor óptimo de x, el cual se calcula de la siguiente manera:

$$x = [A^{\mathbf{T}}\mathrm{W}^{\mathbf{T}}\mathrm{WA}]^{-1}A^{\mathbf{T}}\mathrm{W}^{\mathbf{T}}\mathrm{Wb} \quad (5)$$

Finalmente, factorizando las ecuaciones (3) y (4), se obtiene el algoritmo IRLS:

$$\|e\|_p = \left(\sum_n |e(n)|^{(p-2)} |e(n)|^2\right)^{1/p} \quad (6)$$

Para realizar la transformación anterior a la estructura de ecuaciones ya descrita, las ponderaciones $w(n)$ son estimadas de la siguiente manera:

$$w(n) = e(n)^{\frac{(p-2)}{2}} \quad (7)$$

III.III. II.II. Funcionamiento mecánico del Algoritmo IRLS

Paso I. En el punto inicial t, los pesos asignados en la ecuación (5) se corresponden con el caso clásico en el análisis de regresión.

Paso II. Se estiman los residuos de la ecuación (2).

Paso III. Si el resultado del paso anterior es la no-correspondencia de los residuos obtenidos con los residuos óptimos, el algoritmo procede bajo el criterio de minimización ya planteado (que es el que obedece para asignar y reasignar los pesos a x) a reestimar los residuos en el punto $t + I$, asignando nuevas ponderaciones a la diagonal principal de la matriz de ponderaciones o pesos W, mediante la utilización de (5), definiendo los nuevos pesos mediante la identidad (6).

Paso IV. El procedimiento descrito anteriormente se repite hasta que finalmente las iteraciones converjan a los valores óptimos de los residuos (correspondientes a la minimización óptima de las distancias).

III.III. II.III. Análisis del funcionamiento del Algoritmo IRLS

Se realizan las ponderaciones a la norma y no a la métrica, puesto que la lógica de los espacios euclidianos es que la métrica es inducida por una norma, entonces la métrica es una consecuencia de la norma y no a la inversa (en los espacios definidos). Lo anterior ofrece adicionalmente ciertas ventajas, como por ejemplo ampliar los tipos de métrica para los cuales es válida la iteración algorítmica (mediante la ramificación de métricas que se pueden obtener con una norma en espacios en que la métrica es inducida por una norma). Ponderar los valores de x implica un re-escalamiento (que captura la intuición geométrica de cambiar las medidas de una determinada figura), lo que hace que decrezcan las distancias de las x_i (asignándole nuevas localizaciones a los valores x_i su nueva posición es el valor resultante de multiplicar el valor x_i por la ponderación w_i), es decir, que decrezca la varianza de x. Reducir la varianza es reducir las distancias respecto de la media, lo que implica que los elementos x_i están cada vez más próximos entre sí; lo anterior es una de las implicaciones lógicas del Teorema Central del Límite, que garantiza en ausencia de error sistemático la convergencia a una distribución de probabilidad normal. El incremento asintótico (progresivo y de largo plazo) de la

proximidad entre los x_i es una consecuencia de que el algoritmo IRLS es, como su nombre lo indica, un algoritmo de modelado por una función iterada de la forma $\{x, f(x), f(f(x)), \dots\}$ y debido al teorema de la aplicación contractiva de Banach, el cual garantiza que un mapeo funcional contractivo (que es el tipo de mapeo realizado con las sucesiones de funciones iteradas) es siempre convergente, por lo que la sucesión de funciones iteradas por las que está compuesto el algoritmo IRLS será, también, siempre convergente, cuando el punto fijo atractivo x_0 exista. Por ello, siempre que exista solución al sistema que puede obtenerse mediante mínimos cuadrados ordinarios (esta solución es el punto fijo atractivo), sea una solución exacta o una solución *sparse*[11], es posible aplicar este algoritmo.

III.III. II.IV. Estadísticos Suficientes

Como señalan (Nelder & Wedderburn, 1972, pág. 374), un caso de especial importancia en la estimación estadística de este modelo ocurre cuando el valor del parámetro θ de la distribución del componente aleatorio y el valor pronosticado Y por el modelo lineal coinciden. En el caso descrito antes, tanto $L = zY - g(Y) + h(z)$ como $\frac{\partial L}{\partial \beta_i} = \alpha(\phi)(z - \mu)x_i$ [con base en $\frac{\partial L}{\partial \theta} = \alpha(\phi)(z - \mu)$], *i.e.*, las ecuaciones de máxima verosimilitud. Estas pueden resumirse entonces como una suma indizada a las observaciones $\sum_k (z - \hat{\mu})\, x_{ik} = 0$.

Por tanto, se obtiene la equivalencia $\sum_k z_k x_{ik} = \sum_k \hat{\mu} x_{ik}$. Para variables independientes cualitativas, esto implica que los totales marginales ajustados[12] con respecto a esa variable será igual a los observados.

De la expresión para L es posible observar que las cantidades $\sum_k z_k x_{ik}$ son conjuntos de estadísticos suficientes (que cumplen las condiciones de máxima

---

[11] Una solución *sparse*, significa que la mayoría de los coeficientes de x son ceros, solo unos pocos son distintos de ceros, lo que implica que a medida la densidad o la masa de probabilidad crecen, el modelo necesita cada vez menos variables independientes para explicar a la variable dependiente; a nivel empírico esto se observa en modelos estadísticos para datos de panel en los que se utiliza en los procesos de optimización estadística únicamente algunas columnas. Una solución *sparse* es deseable puesto que con ella se podría evitar un sobreajuste del modelo estadístico en cuanto obliga al algoritmo IRLS a que optimice con menos variables y aún así lograr el resultado óptimo o cercano al óptimo.

[12] Como se señala en (Wikipedia, 2021), el *procedimiento de ajuste iterativo* (conocido también como *ajuste biproporcional* o *biproporción* en Estadística, *algoritmo RAS* en Economía en el contexto del análisis insumo-producto, *raking* en Encuestas por Muestreo y escalamiento matricial en Ciencias de la Computación) es una operación para encontrar la matriz ajustada X (*i.e.*, la matriz del conjunto de datos pronosticados) que es la más cercana a la matriz inicial Z, pero con los totales fila y los totales columna de una matriz objetivo Y (lo que proporciona las restricciones al problema de optimización; el interior de la matriz Y es desconocido. La matriz ajustada de la forma $X = PZQ$, en donde P y Q son matrices diagonales tal que X tiene los márgenes (*i.e.*, los totales por filas y columnas) de Y.

verosimilitud establecidas). Además, en $\frac{\partial L}{\partial \beta_i} = \alpha(\phi)\left\{\frac{\left(\frac{d\mu}{dY}\right)}{V+(z-\mu)\left(\frac{d^2\theta}{dY^2}\right)}\right\}$, $\frac{d^2\theta}{dY^2} = 0$, por lo que $\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} = E\left(\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}\right) = -\alpha(\phi)\left(\frac{\left(\frac{d\mu}{dY}\right)}{V}\right)x_i x_j$.

Cuando θ es también la media de la distribución, *i.e.*, $\mu = \theta = Y$, se tiene entonces el modelo lineal usual con errores normales, que para $g'(\theta) = \theta$ se obtiene el resultado $g(\theta) = \frac{1}{2}\theta^2 + constante$ que determina de forma única la distribución como normal de varianza $\frac{1}{\alpha(\phi)}$ como resultado del siguiente teorema:

(Patil & Shorrock, 1965, pág. 94), *teorema 1*: Es conocido que entre todas las familias del tipo exponencial la forma funcional de la función generadora $f(\omega)$ (una forma de codificar una sucesión infinita de elementos a_n tratándolos como si fuesen coeficientes de una serie de potencias formal -*i.e.*, sin consideraciones sobre convergencia-) caracteriza a la familia. Generalizando la afirmación anterior, sea S una sucesión con un punto límite en Ω_v, en donde v es una función no decreciente real evaluada que sirve como medida del espacio Ω. Si $\mu(\omega)$ se encuentra dentro de S, entonces la familia está determinada dentro de todas las familias exponenciales posibles.

Finalmente, la subclase de modelos para los cuales existen estadísticos suficientes fue descubierto por Cox (1968), mientras que Dempster (1971) extendió los resultados de Cox para incluir varias variables dependientes.

IV. Referencias

Allen, M. (2017). *The SAGE Encyclopedia of COMMUNICATION RESEARCH METHODS.* London: SAGE Publications, Inc.

AMERICAN PSYCHOLOGICAL ASSOCIATION. (15 de Julio de 2021). *level*. Obtenido de APA Dictionary of Pyschology: https://dictionary.apa.org/level

AMERICAN PYSCHOLOGICAL ASSOCIATION. (15 de Julio de 2021). *factor*. Obtenido de APA Dictionary of Pyschology: https://dictionary.apa.org/factor

AMERICAN PYSCHOLOGY ASSOCIATION. (15 de Julio de 2021). *logistic regression (LR)*. Obtenido de APA Dictionary of Pyschology: https://dictionary.apa.org/logistic-regression

Bhuptani, R. (13 de Julio de 2020). *Quora*. Obtenido de What is the difference between linear regression and least squares?:

https://www.quora.com/What-is-the-difference-between-linear-regression-and-least-squares

Burrus, C. S. (7 de Julio de 2021). *Iterative Reweighted Least Squares.* Obtenido de https://cnx.org/exports/92b90377-2b34-49e4-b26f-7fe572db78a1@12.pdf/iterative-reweighted-least-squares-12.pdf

Centro Centroamericano de Población. (28 de Abril de 2021). *Variables y escalas de medición*. Obtenido de Universidad de Costa Rica: https://ccp.ucr.ac.cr/cursos/epidistancia/contenido/2_escmed.html

Greene, W. (2012). *Econometric Analysis* (Séptima ed.). Harlow, Essex, England: Pearson Education Limited.

Gujarati, D., & Porter, D. (8 de Julio de 2010). *Econometría* (Quinta ed.). México, D.F.: McGrawHill Educación. Obtenido de Homocedasticidad.

Haskett, D. R. (10 de Octubre de 2014). "*Mitochondrial DNA and Human Evolution" (1987), by "Mitochondrial DNA and Human Evolution" (1987), by Rebecca Louise Cann, Mark Stoneking, and Allan Charles Wilson*. Obtenido de The Embryo Project Encyclopedia: https://embryo.asu.edu/pages/mitochondrial-dna-and-human-evolution-1987-rebecca-louise-cann-mark-stoneking-and-allan

Kolmogórov, A. N., & Fomin, S. V. (1978). *Elementos de la Teoría de Funciones y del Análisis Funcional* (Tercera ed.). (q. e.-m. Traducido del ruso por Carlos Vega, Trad.) Moscú: MIR.

Lipschutz, S. (1992). *Álgebra Lineal.* Madrid: McGraw-Hill.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (Segunda ed.). London: Chapman and Hall.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135*(3), 370-384.

Online Stat Book. (15 de Julio de 2021). *Levels of an Independent Variable*. Obtenido de Independent and dependent variables: https://onlinestatbook.com/2/introduction/variables.html

Patil, G. P., & Shorrock, R. (1965). On Certain Properties of the Exponential-type Families. *Journal of the Royal Statistical, 27*(1), 94-99.

Perry, J. (2 de Abril de 2014). *NORM TO/FROM METRIC*. Obtenido de The University of Southern Mississippi: https://www.math.usm.edu/perry/old_classes/mat681sp14/norm_and_metric.pdf

Ritchey, F. (2002). *ESTADÍSTICA PARA LAS CIENCIAS SOCIALES. El potencial de la imaginación estadística.* México, D.F.: McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.

StackExchange Cross Validated. (2 de Febrero de 2017). *"Least Squares" and "Linear Regression", are they synonyms?* Obtenido de What is the difference between least squares and linear regression? Is it the same thing?: https://stats.stackexchange.com/questions/259525/least-squares-and-linear-regression-are-they-synonyms

TalkStats. (29 de Noviembre de 2011). *SPSS*. Obtenido de Forums: http://www.talkstats.com/threads/what-is-the-difference-between-a-factor-and-a-covariate-for-multinomial-logistic-reg.21864/

van den Berg, R. G. (15 de Julio de 2021). *Measurement Levels – What and Why?* Obtenido de SPSS Tutorials: https://www.spss-tutorials.com/measurement-levels/

Wikipedia. (21 de Mayo de 2021). *Iterative proportional fitting*. Obtenido de Statistical algorithms: https://en.wikipedia.org/wiki/Iterative_proportional_fitting

Wikipedia. (25 de Febrero de 2021). *Iteratively reweighted least squares*. Obtenido de Least squares: https://en.wikipedia.org/wiki/Iteratively_reweighted_least_squares

Wikipedia. (8 de Julio de 2021). *Lp space*. Obtenido de Measure theory: https://www.wikiwand.com/en/Lp_space