Revista Colombiana de Estadística
Nº 17-18, 1988

EL ANALISIS DE REGRESION
PERSPECTIVA HISTORICA

J. Humberto Mayorga A. y Oscar F. Soto B.

Profesores Asistentes. Departamento de
Matemáticas y Estadística
Universidad Nacional

Resumen. Este artículo presenta un esbozo histórico del desarrollo de la regresión desde Legendre hasta los recientes logros que se han alcanzado en esta área de la Estadística, pretendiendo dar un marco general de referencia sobre la temática.

Se divide la historia en tres eras, para señalar principalmente algunos aspectos de los trabajos de Legendre y Gauss como pioneros en las ideas de regresión y destacar los avances más relevantes de la época actual, señalando algunos de los más renombrados investigadores de cada tema.

Introducción. Tratar aspectos de la historia de la Estadística, es una tarea poco usual en nuestro ambito académico, pero al abordar tal labor, de por si difícil, resulta interesante y apasionante. En verdad, el indagar por el origen, por el nacimiento, por aquella primera oportunidad en que se utilizó un procedimiento estadístico, resultan actividades muy gratas, pues despiertan una curiosidad insaciable por encontrar ese secreto que invitó a las mentes de grandes hombres a crear

ideas estructurales de esta rama del saber humano.

Las más antiguas disciplinas parece hubieran hecho un pacto especial, para contribuir (casi todas), cada una desde su óptica, a la "gestación, nacimiento y crianza" de unas unidades, que luego se asociarían para constituir ese cuerpo conceptual llamado *Estadística*.

Así, la Economía y la Meteorología por su parte, fueron la fuente del Análisis de Series de Tiempo; la Biología admite la cuantificación y dá origen al Análisis de Correlación; el Análisis de Frecuencias, son su popular distribución Ji-cuadrado, se deriva a partir de la constitución de la sociedad humana en otro sujeto de Análisis Estadístico. En fin, la idea de regularidad en el comportamiento de los eventos aleatorios y en las mediciones en los fenómenos naturales, llamó a la puerta de muchas disciplinas y actividades humanas.

En forma similar, el cautivante mundo de la Astronomía con sus abundantes registros y el rigor en su consignación, también fue visitado por esa idea pilar de la regularidad estadística, en momentos en los cuales a nuestros antepasados les asistía el espíritu de independencia de la corona española, dando paso a un punto de vista de análisis con presupuesto propio, que hoy llamamos *Análisis de Regresión*, distinto al también reciente *Teoría de Errores* y que en justicia debería contener en su denominación el apellido del mas destacado de sus pioneros Karl Friedrich Gauss.

El triunvirato de Gauss en Alemania, Legendre en Francia y el menos nombrado Adrain en América, inicia la primera era de la historia del desarrollo de los conceptos, métodos y usos del Análisis de Regresión, la cual cubre el período que se

inicia en 1805 y que termina alrededor de 1935, año en el cual Fisher publica se texto de diseño experimental, y simultaneamente Aitken formula la teoría de los mínimos cuadrados en forma matricial.

La segunda época se caracteriza por el desarrollo conceptual, impulsado por el hecho de producirse en este periodo, en la Teoría de la Estadística un fulgurante avance, que contribuye con la consolidación de la estructura de la Teoría de la Regresión.

Alrededor de 1960, como consecuencia del desarrollo tecnológico en el computo electrónico, comienza una tercera época en la historia de la regresión, en la cual el interés se centra en el aspecto metodológico y los resultados de estos procesos, son rápidamnente implementados en el "soft-ware" de los sistemas de computación.

Era Inicial: Origen del Modelo Lineal

En el año 1805, Adrian Legendre, expone las ideas iniciales del método de mínimos cuadrados, en el apéndice de su trabajo "Nouvelles Méthodes pour la Determination des Orbites des Comètes" en el cual trata sobre "la méthode des moindre carrés", sin ofrecer alguna prueba del método.

Al siguiente año, Gauss reclama el derecho de haber utilizado desde hace 12 años, ese método que Legendre llamó de los mínimos cuadrados y se compromete a presentar sus resultados posteriormente.

Dos años después, R. Adrain, aparentemente sin conocer el

trabajo de Legendre y del aún no publicado método de Gauss, desarrolló el método de los mínimos cuadrados y lo utilizó para resolver varios problemas.

Sin entrar a determinar quien fue realmente el iniciador del método y atendiendo solo a las referencias cronológicas de las publicaciones básicas, se presentan a continuación los principales aportes de los precursores de la regresión.

En su citado trabajo, Legendre afirmaba que en problemas en los cuales se requiere obtener las mas exactas conclusiones a partir de observaciones numéricas, casi siempre se termina por establecer un sistema de ecuaciones de la forma:

$$e_i = \sum_{j=\ell}^{q} \beta_j z_{ji} - x_i \quad (i = 1,2,\ldots,N; N > q)$$

Donde z_{ji}: "*coeficientes conocidos*"

x_i : "*la observación numérica*"

β_j : "*los desconocidos*"

e_i : "*los errores*"

El objetivo era "determinar" (estimar?) los q "desconocidos" (parámetros?), de tal forma que cada "error" (residual?) fuese "muy pequeño" y "los errores extremos", sin considerar el signo, estuviesen "dentro de límites muy cercanos".

El "principio" que Legendre propuso para tal propósito fue la minimización, por variacion en los β, de la suma de cuadrados de los errores, procediendo luego a obtener las ecuaciones normales.

En 1809 Gauss, en su trabajo "Theoria Motus Corporum coelestium in sectionibus conicis solem ambiemtium", cumple con

el compromiso de publicar sus resultados, ubicándose en otra perspectiva, al considerar explícitamente la distribución probabilística de los errores del modelo presentado por Legendre.

El aspecto fundamental de este trabajo de Gauss, esta relacionado con el hecho de haber encontrado que la distribución de los errores, asumida continua, era necesariamente normal, cuando aquellos eran muestreados independientemente de un universo normal de media cero y varianza fija o como Gauss preferia, con precisión fija h, donde $2\sigma^2 = h^{-2}$.

En términos generales el desarrollo de Gauss, fue el siguiente: bajo el supesto de que los errores observacionales son variables aleatorias independientes, con distribución de probabilidad $\phi(e)$ y que los β tienen a-priori distribución uniforme, la distribución a posteriori de los mismos, condicionada a las observaciones, es según la teoría Bayesiana, proporcional a la distribución conjunta de los errores. Como estimador de los β, Gauss toma la moda de esta distribución.

Cuando existe un solo parámetro dicho valor mas probable, corresponde a la media aritmética de los errores, la distribución conjunta de estos es necesariamente normal y la distribución de los β es entonces proporcional a la función de densidad de una normal multivariante. Esta probabilidad se maximiza cuando se minimiza la suma de los cuadrados de los residuales, lo cual corresponde exactamente al principio de los mínimos cuadrados. En 1810, Gauss complementa este trabajo con la presentación de un algoritmo computacional del proceso en su "Disquisitio Palladis".

Posteriormente, entre 1821 y 1826, presenta varios escritos bajo el nombre genérico de "Theoria Combinationis Obser-

vationum erroribus minimis obnoxiae", con los cuales culmina sus aportes al modelo. Caben destacar de este conjunto de trabajos los siguientes hechos:

a. El abandono del supuesto de normalidad, al descubrir una desigualdad general aplicable a cualquier distribución de probabilidad continua unimodal y simétrica alrededor de la moda.

b. La determinación de los estimadores insesgados de los β, que además resultan ser de mínima varianza.

c. La extensión del método, para combinación lineal de los β y la derivación de la varianza del estimador, obteniendose lo que hoy en día se llaman covarianzas entre dos β estimados.

d. La obtención de la expresión para determinar la suma de cuadrados residuales a partir de los β estimados.

e. La presentación de un procedimiento para incluir otros β al modelo, sin tener que recalcular los β ya estimados.

f. La generación de una forma aproximada para estimar la desviación estándar de los errores y la deducción del error estándar de tal estimador.

Aunque Gauss desarrollo los métodos para calcular el error estándar del estimador de cualquier β, parece no haberse referido al problema de decidir si uno o más términos del modelo lineal deberían ser legítimamente descartados por no significantes. Este problema fue considerado por Augustin Louis Cauchy, pero sin referencia alguna a la distribución probabilística de los errores residuales. En este sentido, el trabajo de Cauchy es una regresión a las ideas menos sofisticadas de Legendre y Cauchy en sus memorias litografiadas en 1835, lo presenta como

un problema de interpolación, acompañado de un dispendioso algoritmo de solución.

Bienaymé, Chebyshev, y Gram entre otros, durante 1853 a 1883, al tratar de simplificar el cálculo numérico de algoritmo de Cauchy, introducen y dan los primeros pasos en la utilización de los procesos de ortogonalización, para algunos problemas del modelo lineal. Sin embargo estos resultados iniciales no se llevaron a la práctica, e inclusive fueron por algún tiempo olvidados, para ser retomados en épocas más recientes por autores como Romanovsky (1925-1927) y por Rao (1949), entre otros.

Además contribuyeron con sus estudios al desarrollo de la regresión en esta época, autores como Helmert quien en 1875 al analizar la distribución de la varianza muestral, bajo normalidad, deduce la Ji-cuadrado, la cual es redescubierta posteriormente por K. Pearson, en 1900 para pruebas de bondad de ajuste y Pizzetti quien en 1889 hace una extensión del proceso de Helmert, que lo lleva a encontrar la distribución de la suma de los cuadrados de los errores.

A mediados del siglo diecinueve, aparece dentro de la Estadística un nuevo concepto, el de la distribución de probabilidad multivariante y más especificamente el de la distribución normal multivariante o ley multinormal, que tendría una gran influencia en el mayor desarrollo de los procesos de regresión y correlación.

August Bravais, considera por primera vez en 1846, la distribución de probabilidad conjunta de dos y tres variables, y dentro de su trabajo utiliza tangencialmente el concepto de correlación entre dos variables, aunque sin preocuparse dema-

siado por la aplicación práctica de sus resultados. Posteriormente Schols (1875), amplia un poco los trabajos para tres variables y discute la aplicación de la normal bivariante.

Pero fue Edgeworth, en 1892 el primero en proveer una formulación completamente general de la normal multivariante, desafortunadamente con una notación engorrosa y difícil de manejar, tanto que él mismo solo pudo aplicarla numéricamente para cuatro variables y solo K. Pearson, cuatro años después, la presentó con una notación mejorada.

El pionero en la utilización de las "palabras" regresión y correlación fue Sir Francis Galton, al estudiar en "Hereditary Genius" en 1869, la herencia de la estatura. Efectivamente encontró que las estaturas de los hijos cuyos padres tenían estaturas fuera de la "mediamía", tendían a regresar a la medianía de la estatura de su propia generación hecho que Galton llamó entonces "regresión o reversión", palabra que infortunadamente se generalizó para referirse al estudio de la relación funcional entre variables, en contra de la misma etimología del vocablo.

En 1877, introdujo una medida de tal regresión, que luego renombra co-relación o la actual correlación, asignándole el símbolo "r", inicial de regresión. Es de anotar que los aportes de Galton fueron fundamentalmente empíricos y la mayor crítica negativa que se le hace, es haber pretendido hacer aritmética "por olfato". Posteriormente, a finales de 1890, Weldon introduce el concepto de "coeficiente de correlación" y sugiere que, por lo menos dentro del campo de los estudios genéticos en los cuales tuvieron origen estos temas, este parámetro deberá ser constante.

Como se anotó anteriormente, fue Karl Pearson en 1896 quien presentó una formulación definitiva de estas investigaciones la cual fue integrada al cuerpo conceptual de la Estadística. Pearson, extendiendo las ideas de Galton y Weldon a "p" variables correlacionadas deriva la "superficie normal multivariada" y encuentra que si las desviaciones con respecto a la correspondiente media de $p-1$ de las variables, toman valores determinados, la distribución condicional de la restante variable es normal univariada alrededor de un valor esperado.

La estructura de tal valor esperado, corresponde a una combinación lineal de las desviaciones determinadas de las otras variables y cuyos coeficientes son los regresores o "coeficientes de regresión"; a tal combinación se le dió el nombre de "*regresion de* X_p *en* $X_1, X_2, \ldots, X_{p-1}$". En la estimación de los parámetros de la expresión utilizó un método que es el actualmente conocido como de Máxima Verosimiltud.

En 1897, Yule mostró que para dos, tres y cuatro variables, la ecuación de Pearson, era la misma que resulta de estimar por el método de mínimos cuadrados, un valor de la variable X_p, por medio de una combinación lineal de valores "dados" de las otras variables. Pearson probó posteriormente (1899), que esta combinación lineal es la que tiene máximo coeficiente de correlación con X_p.

Karl Pearson amplió la aplicación del modelo lineal de Gauss a una clase más extensa de problemas que la relativa a la de medición de errores, y por ende permitió una interpretación más general que la del modelo de Legendre.

Con Ronald A. Fisher, culmina esta primera etapa del desarrollo de la regresión, recibiendo de este científico, espe-

cialmente entre 1922 y 1935, dos de los mayores aportes para la teoría del muestreo en regresión, especificamente a nivel de la inferencia. Tales aportes fueron la formulación y la presentación del proceso para pruebas de hipótesis relativas a los coeficientes de regresión, y el desarrollo del análisis de varianza, con la consiguiente deducción de la distribución F y la utilización de la recién deducida distribución de W. Gosset.

Como es de común decir, R. Fisher "no necesita presentación" y realmente tratar de presentar exahustivamente todas sus contribuciones, no solo al desarrollo de la regresión sino de la Estadística en general, se sale de las posibilidades de esta breve exposición, mereciendo quizás un espacio aparte.

Terminamos en este punto el somero esbozo acerca del origen de este proceso estadístico, para dar paso a continuación a la presentación muy esquemática de la época moderna, de una forma un poco diferente a la anterior, debido a la gran cantidad de estudiosos que han contribuido al avance de la misma, lo que impide hacer una pormenorizada referencia a los logros de cada uno de ellos, limitándonos a identificar algunos nombres asociados con los principales temas.

Pasando por alto un cuarto de siglo, en el cual, por el mismo desarrollo de la Teoría Estadística se consolidan y refinan los aspectos teóricos de la regresión, pero con una metodología relativamente incipiente referiremos la que llamamos era metodológica que cubre los años desde 1960 hasta nuestros días, la cual consideramos por su naturaleza más interesante de comentar.

Era Moderna: Desarrollo Metodológico.

La complejidad de los procesos de cálculo y la carencia de herramientas de computo hasta finales de la década de los cincuenta impidieron un mayor desarrollo de la metodología del análisis de regresión y el avance de varias ideas, algunas de ellas manifiestas en publicaciones del siglo pasado, que permanecieron latentes hasta encontrar el auxilio del computador para permitir su implementación. Por esta razón conceptos corrientes hoy día como robustez, colinealidad, selección de variables, entre otros no pertenecían al conjunto de términos de la regresión o solo fueron citados por aquellos años, de una manera muy timida.

Este último período de unos treinta años y que aún no termina, es una era caracterizada por los avances significativos en la metodología de la regresión. Rico en publicaciones y con un elevado número de nombres asociados a sus avances tanto en la estructuración de los conceptos, como en la generación de los procedimientos para su aplicación, logrando de esta manera hacer del análisis de la regresión una herramienta accesible y que con el soporte de la computación, la convierte en una de las más importantes áreas del análisis estadístico.

El preguntar por la adecuación de los supestos de normalidad, independencia y homocedasticidad, buscando la construcción adecuada de un modelo, puede tener respuesta en el *Examen de Residuales*, aspecto que ha interesado a varios autores. Tuckey, Anscombe, Pasternack, Luizzi, Behnken, Snee y Draper, trazan las pautas que permiten este examen, con sus publicaciones entre 1961 y 1972 principalmente.

En el mismo sentido del cumplimiento de los supuestos del

modelo, Box hacia el año 1962 conjuntamente con Tidwell y posteriormente con Cox, impulsan la idea y proponen los métodos para la trasformación de variables. Se destacan también en este tema los aportes de Atkinson, desde 1973, e igualmente los artículos de Cook y Weisbergen en 1982.

La vieja idea de *Selección de Variables*, que se puede derivar indirectamente del trabajo de Cauchy, presentada por él en forma no estadística, pero con una raíz común, es entendida hoy en día como la búsqueda del "mejor modelo" y surge de la pregunta sobre el efecto de añadir o excluir variables en el modelo de regresión. En 1960 aparecen ideas al respecto con el trabajo de Efroymson, en el cual se hace referencia al proceso "Stepwise". Los sentidos de "Forward" y "Backward" de este proceso, fueron analizados 10 años después por Mantel, quien expresa razones de preferencia por el sentido Backward. Los detalles del proceso y el uso de la estadística F como eje del mismo son comentados por Pope y Webster en 1972.

Alrededor de Hocking hay un grupo de autores como Leslie, Pendlenton y La Motte, quienes han publicado varios artículos referentes a la selección de variables, que incluyen los procedimientos computacionales para llevarla a cabo. Es conveniente anotar que este es uno de los temas de regresión, sobre el cual la publicación ha sido bastante prolija y fructífera.

La Colinealidad, es otro concepto que impulsa el desarrollo de procedimientos y criterios para analizarla y/o removerla. Ejemplo de ello es la utilización de la *Regresión Ridge*, como alternativa al método de mínimos cuadrados ordinarios, propuesta por Hoerl y Kennard en 1970 (con orígenes en el año 1962 e implementada computacionalmente en 1981), convirtiéndose en otra de las áreas de la regresión, más trabajadas y por tanto

de gran producción bibliográfica.

El estudio de los residuales, dío origen a su vez en 1973, a la aplicación en regresión de la *Robustez*, campo de la teoría de la estimación estadística recientemente originada y en pleno desarrollo, aplicable a los problemas que surgen por la no aceptación del supuesto de normalidad en la distribución de los errores. Con el trabajo en *Regresión Robusta* de Huber presentado en el citado año y con la publicación de Andrews en el año siguiente se cuenta con un conjunto de estimadores robustos, cuyos procedimientos computacionales son analizados por Holland, Welsch, Denby, Mallows y Larson alrededor de 1977.

Con el termino *Diagnósticos en Regresión*, se distinguen varios procedimientos que tienden a detectar "valores anómalos", como son los "outliers", "casos influenciales", "valores extremos" que de una u otra manera afectan las estimaciones del modelo, con el fin de determinar si un valor o un conjunto de valores tales, debe ser removido, modificado o retenido dentro del mismo. Algunos autores como Hoalglin, Welsch, Andrews, Pregibon, Weiberg y muy especialmente Cook, han ideado a partir de 1977, medidas de diagnóstico que permiten detectar casos individuales o grupos de casos que puedan diferir de la generalidad. Draper y John (1981) discuten los méritos relativos de algunos de estos métodos.

El "problema de enmascaramiento", en el cual la combinación de dos o más casos influenciales, pueden dar un diagnóstico aceptable, en tanto que uno solo no, genera la consideración de medidas de diagnóstico para subconjuntos de casos. Los conceptos de las medidas para estas situaciones, se pueden extender fácilmente, pero como lo anotan Andrews y Pregibon (1978), los cálculos son extensos, por lo cual ha sido necesa-

rio utilizar soluciones gráficas aproximadas, como las sugeridas por Larsen y McCleary en 1972 o por Belsley, Kuh y Welsch en 1980 y resumidas por Mallows en 1982.

Otros temas relativos a la regresión, tales como la regresión no lineal con toda su enorme complejidad especialmente en los procesos infrenciales, la regresión con restricciones, y la muy reciente idea de cointegración, han tenido y siguen teniendo un desarrollo notable en esta época y son muchos los autores contemporaneos, varios de ellos en nuestro país, que continúan cuestionando y ampliando estos procesos.

Epílogo.

Es deseo de los autores, que estas breves notas históricas referentes primordialmente a la Teoría y Métodos de la Regresión, hayan logrado el fin principal de motivar a los lectores a mirar con mayor interes el estudio de la perspectiva histórica, no solo de la Regresión, sino de la Estadística en general, pues el conocer el pasado es fundamental para entender el presente y construir el futuro, verdad que tiene cabida en todos los hechos del ser humano y en particular en el caso de la Estadística, algunas preguntas sobre diversos aspectos solo tienen respuesta por el conocimiento del desarrollo del pensamiento estadístico a través del tiempo.

* *

BIBLIOGRAFIA

Drapper, N.R. y Smith, H. (1981). Applied Regression Analysis (2da. Ed.). John Wiley, New York.

Fisher, B.J. 1978. R.A. Fisher, the life of a scientist. John Wiley and Sons. New York.

Hocking, R.R. 1983. Developments in linear regression methodology: 1956/1982. Technometrics, Vol. 25 N°- 3, Agosto.

Kotz, S.; Johnson, N.L. 1982. Encyclopedia of Statistical Sciences. John Wiley and Sons.

Pearson, E.S., Kendall, M.G. 1970. Studies in the hostory of statistics and probability. Hafner Publishing Company.

*