denarius
revista de economía y administración

Panorama de algunos modelos estadísticos

*Hortensia Moreno Macías**

Resumen

Los modelos estadísticos más conocidos son los modelos de regresión lineal. Está claro que no son los únicos. En este artículo se trata de ofrecer una visión muy general sobre la naturaleza de algunos de los modelos estadísticos que tienen mayor aplicación en fenómenos tanto naturales como sociales, y sobre el proceso de decisión relacionado con estas áreas. Se describen las características más importantes de cada uno de ellos, con la finalidad de que el lector identifique sus diferencias y determine en qué condiciones es conveniente aplicar alguno de ellos para una toma de decisiones exitosa.

Palabras clave: modelos estadísticos, procesos de toma de decisiones.

1. Modelos matemáticos y estadísticos

De acuerdo con algunos autores, un modelo es una representación abstracta que trata de explicar, de la mejor manera posible, un fenómeno o proceso de la vida real. Hoy día es común hablar de modelos en diversas áreas del conocimiento:

* Profesor investigador del Departamento de Economía, División de Ciencias Sociales y Humanidades, UAM-I.

modelos biológicos, físicos, económicos y sociales, entre otros. Los modelos científicos separan un sistema en sus partes elementales y abstraen las relaciones existentes entre ellas para proporcionar una estructura lógica que permita su estudio de manera objetiva y a partir de la cual se obtengan inferencias. Las matemáticas forman un sistema lógico muy poderoso, de tal suerte que la mayoría de los modelos se desarrollan como abstracciones matemáticas de fenómenos reales. Un modelo puramente matemático es un recurso determinista en el que, dado un conjunto de puntos iniciales, éste predice con absoluta certeza el resultado. Pueden mencionarse un gran número de ejemplos de modelos matemáticos; por su parte, Carroll y Ruppert (1988) citan un modelo para determinar el número de peces nuevos (R) en un cardumen en función del número de peces originalmente observados (S). La expresión matemática es:

$$R = \beta_1 S \exp(-\beta_2 S)$$

con β_1 y $\beta_2 \geq 0$ como parámetros del modelo.

Carroll y Ruppert analizan la pertinencia de explicar tal fenómeno a través de un modelo determinista, argumentando que no es posible controlar todos aquellos factores que participan de manera implícita o explícita en el proceso de reproducción de los peces. De hecho, el modelo no toma en consideración factores ambientales importantes, como pueden ser la fuerza de las corrientes marinas, los niveles de contaminación, la presencia de depredadores y la disponibilidad de alimento. Por otro lado, los errores de medición (en este caso, la estrategia del conteo de peces) también provocan cierto alejamiento de los resultados con respecto al modelo determinista, por lo que es de gran importancia no perderlos de vista.

Es así como la presencia de factores no conocidos o de naturaleza no controlable motivan la necesidad de incluir en el modelo un "término aleatorio", surgiendo así los modelos estadísticos.

Schabenberger y Pierce (2000) citan un modelo matemático para determinar la concentración (±) de cierto contaminante en un río en el punto s y en el tiempo t (Beltrami, 1988):

$$\alpha(s,t) = \alpha_0 (s - ct) \exp(-\mu t)$$

donde:

μ es una constante de proporcionalidad que mide la eficiencia de la descomposición bacterial del contaminante.

c es la velocidad del agua.

$\alpha_0(s)$ es la concentración inicial del contaminante en el sitio s.

La incertidumbre del efecto de una localización particular a lo largo del río y el punto en el tiempo provocaron que tales autores modificaran el modelo para dar paso a la siguiente expresión:

$$\alpha(s,t) = \alpha_0(s - ct)\exp(-\mu t) + \varepsilon$$

donde:

ε es una variable aleatoria con media cero, varianza σ^2 y alguna distribución de probabilidad.

Se obtiene así un modelo estadístico.

De acuerdo con Shabenberger y Pierce (2002), dentro de las razones que justifican el uso de un modelo estadístico en lugar de uno matemático se pueden mencionar las siguientes:

- El modelo no es correcto para una observación en particular, pero es correcto en promedio para el conjunto de observaciones.
- Es necesario hacer supuestos y omisiones para lograr la abstracción del fenómeno.
- Aunque se conozca la naturaleza de todos los factores que influyen en el fenómeno, puede ser imposible medir o controlar todas las variables.

Y, en cuanto a los aspectos funcionales de los modelos estadísticos, se puede decir que:

- Con frecuencia un modelo estadístico es más parsimonioso que uno matemático.

- Los modelos estadísticos describen las propiedades distribucionales de una o más variables respuesta a través de la descomposición de la variabilidad en fuentes conocidas y desconocidas.

- Los modelos estadísticos se suponen correctos en promedio. La calidad del modelo no está en función de su complejidad o tamaño, pero sí es determinada a través de un diagnóstico y su utilidad para contestar las preguntas que lo motivaron.

2. Población y grado de generalidad

Los conceptos de población y de grado de generalidad de ésta son fundamentales en la tarea del modelado estadístico.

Está claro que una población es un conjunto de elementos con algunas características comunes y otras que varían de un elemento a otro. Por ejemplo, al hablar del grupo de mujeres mexicanas en edad reproductiva se tiene que las características comunes son: la nacionalidad, el género y el grupo de edad (reproductiva). Las características no comunes pueden ser: el lugar de residencia en la República Mexicana, la pertenencia o no a una etnia indígena, el acceso a los servicios públicos de salud o la carencia de ellos, y los diversos niveles socioeconómicos, por mencionar sólo algunos.

La media de la variable respuesta depende de los factores comunes, mientras que su varianza depende de los no comunes. Esto es, la varianza será más pequeña conforme la población sea menos general, porque depende de menos factores no comunes.

En la investigación comparativa se busca valorar el cambio en las medias de varias poblaciones con el mismo grado de generalidad y determinar, de la manera más objetiva posible, si las diferencias son atribuibles a los factores estudiados (condiciones constantes dentro de las poblaciones, pero que varían entre las poblaciones estudiadas) o sólo son un reflejo de la variabilidad aleatoria atribuible al fenómeno.

3. Tipos de modelos estadísticos

En términos generales, los modelos estadísticos están constituidos por los siguientes componentes:

- Variable respuesta o valores observados (Y)
- Parámetros $(\beta_1,\beta_2,\ldots,\beta_p)$
- Parte sistemática $(\mu(X_1, X_2,..., X_p))$
- Parte aleatoria (ε)

Esto es, viéndolo en forma de ecuación, se tiene que:

Variable respuesta = parte sistemática + parte aleatoria

$$Y = \mu(X_1, X_2,..., X_p) + \varepsilon$$

y los parámetros $\beta_1,\beta_2,\ldots, \beta_p$ son los coeficientes[1] de cada uno de los términos de la parte sistemática.

Nótese que se estima el valor promedio de la variable respuesta, es decir, se busca un modelo con la expresión:

[1] Desde el enfoque de la estadística clásica, los parámetros son valores fijos pero desconocidos.

$$E(Y) = \mu(X_1, X_2, ..., X_p) = \sum_{j=1}^{p} \beta_j x_j$$

La forma de los diferentes componentes es la que determina el tipo de modelo estadístico, es decir, el modelo para una variable respuesta continua es diferente del que se aplica en el caso en que la respuesta es categórica; la forma de los parámetros permite clasificar los modelos en lineales o no lineales.

A continuación se presenta una breve caracterización de cada uno de los diferentes tipos de algunos de los modelos estadísticos más utilizados en el área social.

3.1 Modelos lineales

Los modelos lineales son los modelos estadísticos más comunes, en particular el modelo de regresión lineal clásico.

La forma del *modelo lineal básico* es:

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

donde :

Y_i es un valor de la variable respuesta o dependiente.

x_i es un valor de la variable explicativa, regresora o independiente.

β_0, β_1 son los parámetros que dan cuenta de la ordenada al origen y la pendiente de la recta que describe, de manera aproximada, la relación entre Y_i y x_i.

ε_i es el error aleatorio. Es un valor no observable. En realidad este término se estudia a través de los residuos.

Se habla de un *modelo de regresión lineal múltiple* cuando se tiene más de una variable independiente. El número de parámetros β_i por estimar es igual al número de variables explicativas más uno (β_0 corresponde a la ordenada al origen).

Por medio de la muestra, se pueden obtener valores estimados para la variable respuesta después de un proceso de estimación de los parámetros; esto es,

$$E(Y_i) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{p_i}$$

Los métodos de estimación más usados son mínimos cuadrados y máxima verosimilitud.

La prueba de hipótesis respecto al valor de los parámetros es:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

Es a través de un modelo de regresión que se evalúa el efecto sistemático de las variables regresoras en la variable respuesta; en este sentido, la hipótesis nula establece que la variable regresora x_j no tiene efecto en la variable dependiente, por lo que estadísticamente no se justifica su presencia en el modelo; sin embargo, bajo un argumento social (o de la naturaleza del fenómeno en estudio) relevante, esta variable puede permanecer en el modelo.

La idea de un modelo "adecuado" puede ser intuitiva si se parte de que se pretende encontrar aquellos factores que participan sistemáticamente en el fenómeno y que lo explican en gran medida, dejando sólo una "pequeña" parte en el término aleatorio. A través de una tabla de análisis de varianza se evalúa de manera objetiva si la porción de la variabilidad del fenómeno explicada por el modelo es estadísticamente mayor que la parte aleatoria.

El coeficiente de determinación es una estadística importante para tomar en cuenta en la valoración de la bondad del ajuste del modelo, es decir, para valorar qué tan bien se ajustan los valores esperados a los observados. Este coeficiente toma valores entre 0 y 1 indicando la proporción de la variación total de la variable respuesta explicada por el modelo. En los resultados por computadora se puede identificar como R^2.

Por otro lado, debido a que todos los modelos estadísticos se basan en supuestos, la verificación de su cumplimiento es una parte importante de la evaluación del modelo.

Los *supuestos básicos del modelo de regresión lineal* son:

- *Aditividad* entre los términos participantes. Esto es, se supone que el modelo de medias como suma de términos es correcto.

- *Normalidad de la distribución de los valores de la variable respuesta al interior de cada población*. Bajo el supuesto de que el modelo de medias es adecuado, los errores en su conjunto siguen una distribución normal con media cero y varianza $\tilde{A}^2$. Si al interior de cada población se considera que los infinitos factores no comunes o fuentes de error son de poca importancia, se espera que se cumpla la normalidad.

- *Homoscedasticidad*. En la medida en que el grado de generalidad de las poblaciones sea el mismo, se espera que las fuentes de error también sean similares, y en ese sentido tendrá sustento el supuesto de homogeneidad de varianzas.

- *Independencia de errores.* El error cometido en una unidad de estudio debe ser independiente del cometido en cualquier otra unidad. Si las unidades forman una muestra aleatoria y los procesos de medición no tienen un efecto sistemático por grupos de unidades, este supuesto se cumple.

El proceso que permite valorar el cumplimiento de los supuestos y, en ese sentido, la pertinencia de aplicar este tipo de modelos se llama diagnóstico.

El término aleatorio (el error) es una pieza fundamental en la valoración de los supuestos. Sin embargo, los errores son variables aleatorias no observables y, por lo tanto, no son sujetos de evaluación. A través de la muestra, y después de elegir un modelo, sólo se pueden calcular los residuos. Ésta es la razón por la que el diagnóstico del modelo se realiza con los residuos.

Los residuos se definen como la diferencia entre los valores realmente observados y los valores esperados a través del modelo, es decir,

$$e_i = Y_i - \hat{Y}_i$$

Los métodos gráficos para elaborar el diagnóstico del modelo constan básicamente de diagramas de dispersión. A continuación se mencionan las gráficas más comunes en el diagnóstico y su interpretación.

La construcción de la gráfica de dispersión entre los residuos (en el eje de las ordenadas) y los valores estimados (en el eje de las abscisas) permite identificar si aún hay factores importantes que se han omitido (se observa una nube de puntos con cierta tendencia). Si el modelo es correcto, se espera observar una nube de puntos con dispersión completamente aleatoria. La figura 1 muestra un ejemplo de diagrama de dispersión en el que los puntos tienen un comportamiento aleatorio, por lo que se puede pensar que no se ha omitido algún factor importante en el modelo. La gráfica pertenece a los resultados obtenidos mediante el paquete estadístico JMP® (2002).

- Es muy útil observar además los diagramas de dispersión entre los residuos y cada una de las variables observadas, tanto las que se han incluido en el modelo como las que faltan por incluir. Esto puede dar una pauta para determinar la pertinencia de incluir más variables en el modelo.

- Un diagrama de dispersión entre los residuos y el orden en que se obtuvieron las observaciones en el tiempo es útil para identificar la presencia de correlación serial y, en consecuencia, la violación al supuesto de independencia.

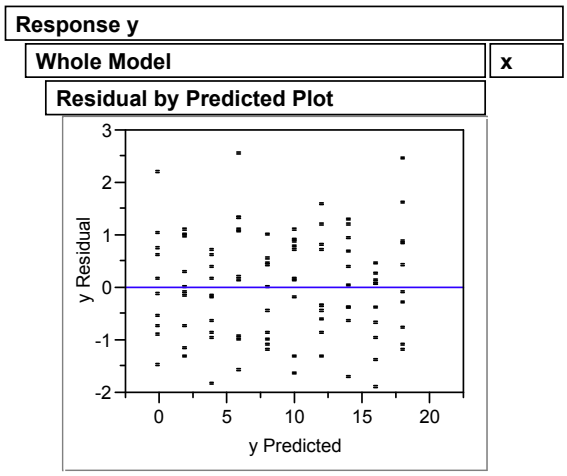


Figura 1. Gráfica de dispersión entre valores ajustados por el modelo y los residuos.

- Si en el diagrama de dispersión entre residuos y valores esperados se observa una forma de "embudo" o de "moño", se tiene evidencia para sospechar la violación al supuesto de homoscedasticidad. La figura 2 muestra un ejemplo de datos con problema de heteroscedasticidad.

- Una gráfica de los residuos en escala normal (bajo la escala de los percentiles de la distribución normal) permite evaluar el supuesto de normalidad en los errores. La gran mayoría de los paquetes estadísticos cuentan con un módulo que realiza dicha gráfica con gran facilidad. Puede aparecer con el nombre de "Normal Probability Plot" o "Normal Quantile Plot". La figura 3 muestra el resultado de la evaluación del supuesto de normalidad con el paquete JMP® (2002). Este despliegue incluye un histograma, un diagrama de caja, la gráfica en escala normal y la prueba de normalidad Shapiro-Wilk. En este caso se puede observar que el supuesto de normalidad es sustentable.

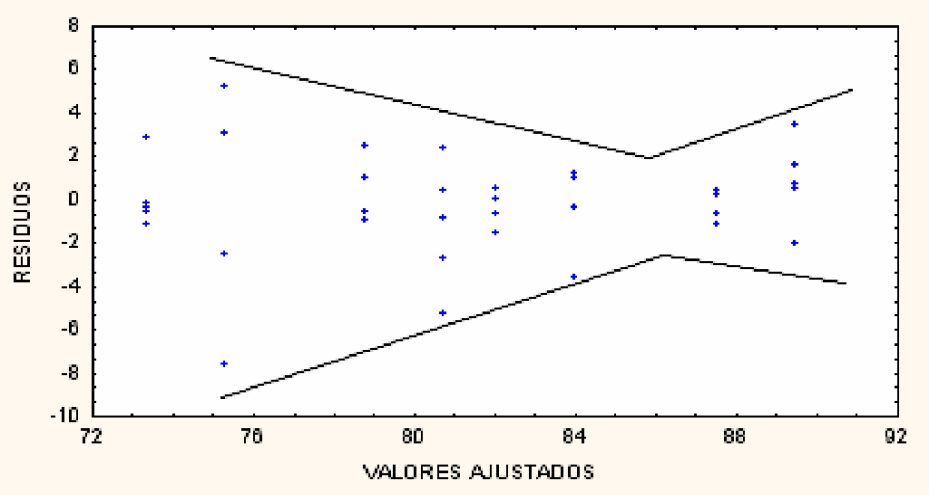


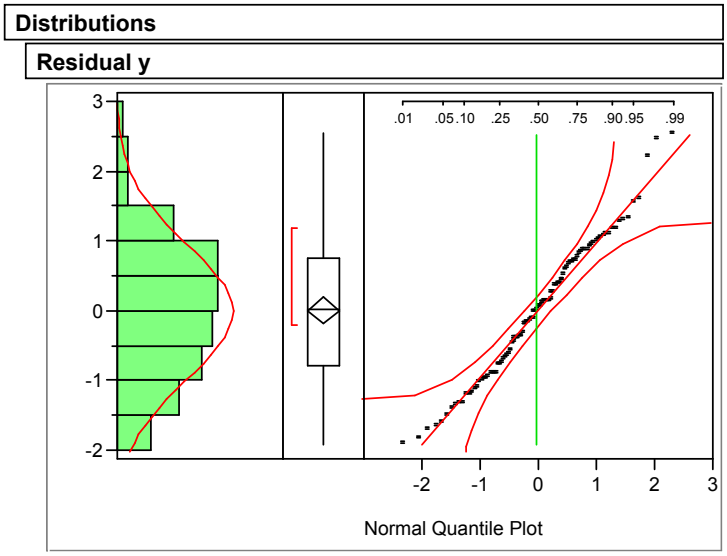
Figura 2. Diagrama de dispersión que muestra forma de "moño", haciendo evidente la violación del supuesto de homoscedasticidad.

Si después de haber realizado el diagnóstico se observa que no se cumple el supuesto de normalidad, con frecuencia se recomienda aplicar alguna transformación[2] a la variable respuesta. Esta alternativa tiene como consecuencia la complejidad de la interpretación de los resultados en términos de la variable transformada.

La violación del supuesto de homoscedasticidad tiene básicamente dos alternativas para corregirla: aplicar una transformación, como en el caso anterior, o usar el método de mínimos cuadrados ponderados en la estimación de los parámetros.

La violación al supuesto de independencia sugiere la aplicación de modelos en presencia de correlación serial (series de tiempo)

[2] Se puede usar el método de Box-Cox para determinar el tipo de transformación que se requiere. Véase Box-Cox (1964).



Moments

Mean	2.576e-15
Std Dev	0.9701766
Std Err Mean	0.0970177
upper 95% Mean	0.1925052
lower 95% Mean	-0.192505
N	100

Fitted Normal

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.971221	0.1689

Figura 3. Resultado de la evaluación del supuesto de normalidad en residuos.

3.2 Modelos no lineales

Con cierta frecuencia, la diferencia entre modelos lineales y no lineales se asocia con la apariencia gráfica de los valores predichos a lo largo del recorrido de una variable explicativa; sin embargo, la no linealidad no se refiere a la curvatura de la función media como una función de covariables. En realidad, se dice que un modelo es no lineal si al menos una de las derivadas de la función media con

respecto a los parámetros depende de al menos un parámetro. Esto es, la linealidad se refiere a linealidad en los parámetros y no en las covariables. Por ejemplo, el polinomio

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^{\,2} + e_i$$

es un modelo lineal aunque al graficar $\hat{y}$ contra x se observe una curva. De hecho, se puede verificar que las derivadas de tal expresión con respecto a cada uno de sus parámetros son funciones que no dependen de los parámetros, esto es:

$$\frac{\partial(\beta_0 + \beta_1 x_i + \beta_2 x_i^{\,2})}{\partial \beta_0} = 1$$

$$\frac{\partial(\beta_0 + \beta_1 x_i + \beta_2 x_i^{\,2})}{\partial \beta_0} = x_i$$

$$\frac{\partial(\beta_0 + \beta_1 x_i + \beta_2 x_i^{\,2})}{\partial \beta_2} = x_i^{\,2}$$

Por otro lado, un ejemplo de modelo no lineal es el siguiente:

$$Y_i = \beta_0 (1 + e^{\beta_1 x_i}) + e_i$$

La no linealidad se puede verificar a través de las derivadas:

$$\frac{\partial(\beta_0 (1 + e^{\beta_1 x_i}))}{\partial \beta_0} = 1 + e^{\beta_1 x_i}$$

$$\frac{\partial(\beta_0 (1 + e^{\beta_1 x_i}))}{\partial \beta_1} = \beta_0 x_i e^{\beta_1 x_i}$$

mismas que dependen de los parámetros del modelo.

Un modelo puede ser lineal en algunos parámetros y no lineal en otros. Por ejemplo,

$$Y_i = \beta_0 + e^{\beta_1 x_i}$$

es un modelo lineal en β_0 y no lineal en β_1. Pero si un modelo es no lineal en al menos un parámetro, el modelo completo se considera no lineal.

De acuerdo con Shabenberger y Pierce (2002), los modelos no lineales tienen las siguientes ventajas sobre los modelos lineales:

- Surgieron de las teorías y principios de fenómenos físicos, químicos y biológicos.
- Este tipo de modelos suelen requerir menos parámetros que los modelos lineales.
- Requieren información sustancial del fenómeno en estudio.

Dentro de las desventajas se mencionan las siguientes:

- El proceso de estimación es más complicado por ser iterativo.
- Se requieren valores iniciales de los parámetros.
- Se necesita un método para decidir cuándo detener las iteraciones.
- Proporcionan sólo inferencias aproximadas, basadas en resultados asintóticos.

Al igual que en los modelos lineales, en los no lineales persiste la idea de minimizar la suma de cuadrados de las desviaciones entre los valores observados y los ajustados por el modelo. Éste es el principio que siguen los algoritmos

iterativos usados para la estimación de los parámetros. Los algoritmos más conocidos son el Gauss-Newton y el Newton-Raphson.

La prueba de hipótesis con respecto al valor de los parámetros es la misma que en el caso de los modelos lineales, esto es:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

y la interpretación de la hipótesis nula es igual que en el caso lineal.

La inferencia estadística en modelos no lineales se basa en la distribución asintótica de los estimadores de los parámetros, por lo que ésta es aproximada.

En cuanto a las aplicaciones, Neter y Wasserman (1989) ejemplifican de manera muy sencilla el uso de modelos no lineales en un problema de administración hospitalaria. En este ejemplo se busca predecir el grado de recuperación a largo plazo de los pacientes que son dados de alta después de haber permanecido hospitalizados. La variable respuesta (Y) es el valor de un índice de diagnóstico asignado a cada paciente y la variable regresora es el número de días de hospitalización. El modelo propuesto es de la forma

$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i$$

Por otro lado, Schabenberger y Pierce (2002) ejemplifican estos modelos con un experimento de fertilización en la caña de azúcar a base de diferentes dosis de nitrógeno. La expresión del modelo propuesto es:

$$Y_i = \alpha(1 - \exp\{-k(x_i - x_0)\}) + e_i$$

En ambos casos, los resultados del ajuste del modelo proporcionan información importante en la toma de decisiones, ya sea para determinar el tiempo promedio adecuado necesario para la recuperación de un paciente (dependiendo de su padecimiento) o para determinar la dosis óptima de nitrógeno en la fertilización de caña de azúcar.

Dada la relevancia de la información que aporta el modelo, el uso de herramientas de diagnóstico para examinar la aptitud del modelo (sea lineal o no lineal) juega un papel muy importante.

El coeficiente de determinación es un valor que debe tomarse en cuenta para evaluar el modelo. La interpretación es la misma que en el caso lineal: conforme se acerca a 1, la proporción de variabilidad explicada por el modelo es mayor.

Los supuestos del modelo son los mismos que en el caso de los modelos lineales, con la única diferencia de que la suma de los residuos no necesariamente es igual a cero. En este sentido, las estrategias gráficas de análisis de residuos son las mismas: diagramas de dispersión entre residuos y valores ajustados para verificar homoscedasticidad o variables regresoras, incluidas o no en el modelo, para identificar tendencias sistemáticas que deben considerarse; y gráficas de los residuos en escala normal para verificar normalidad y de residuos contra el orden en que se tomaron las observaciones para verificar la independencia en el tiempo.

Las medidas remediales ante la violación de los supuestos básicos del modelo también se basan en transformaciones para recuperar normalidad o para estabilizar varianzas.

3.3 Modelos lineales generalizados

Los modelos lineales generalizados son modelos estadísticos para una amplia familia de distribuciones de probabilidad como son la gama, beta, binomial y Poisson, entre otras. De hecho, el modelo de regresión clásico con el supuesto de normalidad en los errores es un caso particular de los modelos lineales generalizados.

En este tipo de modelos se supone independencia y varianza constante en los errores. La distribución de los errores puede ser cualquiera que pertenezca a la familia exponencial.[3]

[3] Familia de funciones de densidad de probabilidad que puede expresarse, en el caso de un parámetro, de la siguiente manera: $f(x;\theta) = a(\theta)b(x)\exp[c(\theta)d(x)]$

Estos modelos relacionan la media de una variable respuesta con una combinación lineal de variables explicativas a través de una función liga (g), es decir:

$$g(E(Y_i)) = g(\hat{Y}_i) = x_i'\beta$$

La función liga transforma los valores esperados de la respuesta en una escala lineal donde los efectos de las variables independientes son aditivos.

La forma de la función liga depende de la naturaleza distribucional de la variable Y. En las secciones posteriores se verá la expresión de g para el caso de la distribución binomial y Poisson, quizá los modelos lineales generalizados más solicitados.

3.3.1 Regresión logística

La principal diferencia entre un modelo de regresión logística y un modelo de regresión lineal es que, en este caso, la variable respuesta es una variable binaria o dicotómica: presencia o ausencia de algún atributo en particular. Esto es, las observaciones son variables aleatorias independientes que siguen una distribución binomial. Esta diferencia se refleja en la selección del modelo paramétrico y sus respectivos supuestos.

Mientras que en el modelo de regresión lineal se buscan aquellos factores que tienen un efecto sistemático en la media condicional de la variable respuesta (Y) dado un valor de la variable regresora ($E(Y/x) = \beta_0 + \beta_1 x$), en el caso de la regresión logística se trata de buscar aquellos factores que tienen un efecto sistemático sobre la probabilidad de pertenecer al grupo que posee el atributo de interés, esto es:

$$E(Y_i) = \pi_i(x) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})}$$

Donde:

π_i es la probabilidad de poseer el atributo.

y $\beta_0, \beta_1, \ldots,^2_p$ son los parámetros del modelo.

De esta manera, es claro que en la regresión lineal el valor promedio de la variable regresora puede tomar valores a lo largo de toda la recta real, mientras que en el caso de la regresión logística este valor está restringido a tomar valores entre 0 y 1.

La función liga en este tipo de modelos se llama logit y se define de la siguiente manera:

$$g(\pi_i(x)) = \log \frac{\pi_i(x)}{1-\pi_i(x)} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

El método de máxima verosimilitud es el más utilizado en la estimación de los parámetros.

La prueba de hipótesis con respecto al valor de los parámetros es la misma que en el caso de los modelos anteriormente expuestos.

El coeficiente de determinación que se usa en los modelos lineales para valorar la bondad del ajuste es reemplazado por una estadística llamada devianza, que mide, más que la proporción de varianza explicada por el modelo, la discrepancia del ajuste.

En cuanto al diagnóstico del modelo, la distribución binomial y no la normal es la que describe el comportamiento de los errores de un modelo de regresión logística, ya que la variable respuesta puede expresarse de la siguiente manera: $y = \pi(x) + \varepsilon$. De aquí que ε pueda tomar sólo uno de dos valores:

- Si y=1, entonces ε =1-$\pi(x)$ con probabilidad $\pi(x)$

- Si y=0, entonces $\varepsilon = -\pi(x)$ con probabilidad $1-\pi(x)$

Por esta razón, el diagnóstico se basa, más que en los residuos, en las probabilidades y en algunas funciones de influencia. Para la comprensión más detallada de este aspecto por demás importante del ajuste de un modelo logístico, es recomendable leer el capítulo 5 del libro de Hosmer Lemeshow (2002).

Como ejemplo de aplicación de los modelos logísticos, se puede mencionar el que desarrolló Menard (1995) para evaluar el efecto del género, la raza y la relación con delincuentes en la probabilidad de ser adicto a la mariguana. Por su parte, Hosmer y Lemeshow (2002) proponen un modelo para la probabilidad de que un neonato presente bajo peso al nacer en función de características biológicas y sociales de la madre.

En ambos casos, la identificación de los factores importantes proporciona información por demás valiosa para la toma de decisiones en políticas de salud de las comunidades correspondeintes

3.3.2 Regresión Poisson

Los modelos para conteos de ocurrencia de eventos aleatorios independientes a lo largo de un periodo de tiempo son los llamados modelos de regresión Poisson. Se encuentran ejemplos en diferentes áreas de estudio, desde el conteo de bacterias en un cultivo hasta el número de accidentes de trabajo en una fábrica o muertes por suicidio en alguna ciudad importante.

El interés de los modelos de regresión Poisson por lo general radica en la estimación de tasas o incidencias de eventos, así como en la determinación de su relación con un conjunto de variables explicativos.

Se supone que $y_1, y_2, ..., y_n$ siguen una distribución Poisson con parámetro μ, donde cada y_i representa el número de eventos que ocurren en un periodo de tiempo.

Si se tiene una sola variable explicativa (x), el modelo de regresión para el promedio de ocurrencias del evento aleatorio por unidad de tiempo (μ) es:

$$\mu = e^{\alpha} e^{\beta x}$$

La función liga g para estos modelos es el logaritmo natural, de tal manera que al aplicarla a la expresión anterior se obtiene un modelo log-ineal

$$\ln(\mu) = \alpha + \beta x$$

Si el interés principal es modelar tasas de ocurrencia, en el modelo se debe incluir un término llamado *offset* y que es el logaritmo del tiempo de exposición. Es decir, si el tiempo de exposición se denota con N, la tasa de ocurrencia es $\frac{Y}{N}$ y el valor esperado se escribe como $\frac{\mu}{N}$.

Al modelar esta tasa con un modelo log-lineal se tiene:

$$\log(\frac{\mu}{N}) = \beta_0 + \beta_1 x$$

por lo que

$$\log \mu = \beta_0 + \beta_1 x + \log(N)$$

El término *log(N)* es el *offset* y debe tomarse en cuenta en el proceso de estimación.

El método de máxima verosimilitud vuelve a ser el método por excelencia para la estimación de los parámetros.

Las hipótesis por probar con respecto al valor de los parámetros del modelo son las mismas que en los casos anteriores.

Como en el caso de la regresión logística, la devianza es una estadística importante para valorar la bondad del ajuste del modelo.

Stokes, Davis y Koch (2000) ilustran este tema por medio de un modelo para estimar la densidad de incidencia de melanoma entre hombres blancos ajustado por grupo de edad y región de residencia. En este ejemplo los autores toman como *offset* el logaritmo del total de personas por grupo, es decir, por región y grupo de edad, mientras que la variable respuesta es el total de personas que presentaron el evento dentro de cada grupo.

Comentarios finales

En este documento se han presentado de manera muy resumida algunos de los modelos estadísticos importantes para estudiar fenómenos sociales, con la idea de proporcionar un panorama de las herramientas estadísticas en el arte de modelar. Se ha usado un lenguaje poco técnico no con la intención de minimizar las expresiones matemáticas formales, sino con el objetivo de lograr el interés de estudiosos de otras áreas que no necesariamente están familiarizados con los métodos cuantitativos. En este principio también se incluyen las expresiones más importantes de los modelos para que el lector tenga un punto de referencia con la literatura formal. La lista de modelos ausentes en este escrito es larga, y por supuesto que la cantidad de material que se queda en el tintero es extensa. Se requerirían volúmenes completos para analizar otros conceptos, detalles, ejemplos, aplicaciones y recomendaciones.

En cuanto al uso de paquetería especializada, también se tiene una larga lista; sin embargo, dentro de los productos más utilizados están SPSS, JMP, STATA y SAS. Aunque las últimas versiones de estos paquetes trabajan bajo la modalidad de ventanas, quizá los más amigables sean los dos primeros.

BIBLIOGRAFÍA

Beltrami, E. *Mathematics for Dynamic Modeling*, 2nd ed., Academic Press, San Diego, 1988, p. 86.

Box, G.E.P. y D.R. Cox. "An Analysis of Transformations", *Journal of the Royal Statistical Society*, B, vol. 26, 1964, pp. 211-243.

Carrol R. y D. Ruppert. *Transformation and Weighting in Regression*, Chapman and Hall, New York, 1988, pp. 1-2.

Dobson, A. *An Introduction to Generalized Linear Models*, Chapman & Hall, New York, 1990.

Harrell, F. *Regression Modeling Strategies*, Springer-Verlag, New York, 2001.

Hosmer D. Jr. y S. Lemeshow. *Applied Logistic Regression*, John Wiley & sons, New York, 2002.

JMP® User´s Guide. The statistical Discovery software, 2003 SAS Institute Inc.

McCullagh, P. y J.A. Nelder. *Generalized Linear Models*, Chapman & Hall, New York, 1984.

Menard, Scott. *Applied Logistic Regression Analysis*, SAGE Publications, New Delhi, 1995.

Méndez, I. y H. Moreno. *Modelos estadísticos en la investigación comparativa*, 2ª ed., (Serie Monografías). IIMAS/UNAM, México, 2003.

Neter, J., W. Wasserman y M. Kutner. *Applied Linear Regression Models*, IRWIN, 1989.

Shabenberger, O. y F. Pierce. *Contemporary Statistical Models*, CRC Press, New York, 2002.

Stokes, M., C. Davis y G. Koch. *Categorical Data Analysis*, SAS Institute, 2000.