Mínimos cuadrados versus verosimilitud

Pedro Cervantes-Hernández*, Andrea Flores-Gómez & Blanca Sánchez-Meraz

Una de las preocupaciones más comunes en la investigación ecológico-biológica, es la obtención de predicciones confiables derivadas de modelos lineales y no lineales. En primera instancia, lo anterior depende en gran medida de la calidad de la información; si esto es corroborado con anterioridad, entonces el error de ajuste (ε) entre la variable dependiente observada (Y_i) y su correspondiente valor esperado ($\hat{Y}_i$), dependerá de la precisión con la que se hayan estimado los parámetros en ambos tipos de modelos.

Los mínimos cuadrados (MC) y la verosimilitud (*L*) corresponden a un grupo de técnicas que se especializan en la estimación de parámetros, siendo los MC, los más ampliamente utilizados. Al emplear los MC como técnica para estimar parámetros en los modelos lineales y no lineales, las predicciones de $\hat{Y}_i$ con respecto de Y_i, consideran un grupo de garantías para validar los residuos entre éstos, de entre los cuales destacan: el coeficiente de correlación (r), el coeficiente de determinación (r^2), el mapa de residuos y el nivel de significancia o valor de *p*. Lo anterior es comúnmente empleado, cuando la técnica de MC es utilizada para generar una estimación puntual por parámetro a estimar.

A diferencia de los MC, la verosimilitud no requiere del grupo de garantías antes mencionadas para validar las predicciones de $\hat{Y}_i$ con respecto de Y_i. La técnica es probabilística y emplea a los MC como auxiliar en la estimación de parámetros, por lo que se considera más precisa que éstos últimos. La verosimilitud está tomando un gran auge en la vida moderna del quehacer científico, por lo cual el objetivo de este trabajo es mostrar su utilidad, aplicación y ventajas en el área de la investigación ecológico-biológica.

En probabilidad clásica, la probabilidad de un evento P(y_i) depende directamente de un espacio muestral conocido (*S*), por lo que cualquier probabilidad y_i en S es conocida con anterioridad. La probabilidad condicional de y_i dado que se conoce *P* es:

$$P\{y_i / P\} \qquad (1)$$

y

$$\sum_{i=1}^{I} P\{y_i / P\} = 1 \qquad (2)$$

En la ecuación 1, el subíndice *i* es asignado a *y*, e indica que y_i puede tomar valores entre [*i, I*]. En la ecuación 2, la suma total de las probabilidades estimadas para y_i en *S* es igual a 1.

En el caso de la verosimilitud, la probabilidad condicional es evaluada de la siguiente manera:

$$L\{P_m / y\} \qquad (3)$$

y

$$\sum_{m=1}^{M} L\{P_m / y\} \qquad (4)$$

En la ecuación 3, el subíndice *m* asignado a *P* indica que P_m puede tomar valores entre *m* y *M*, designando al total de probabilidades en las cuales *y* puede ocurrir. En la ecuación 4, la suma de las probabilidades es ≠ 1 ya que el espacio muestral *S* es desconocido.

Suponiendo que y_i y P_m en las ecuaciones 1 y 3 son ahora el parámetro a estimar (θ), las

Universidad del Mar, campus Puerto Ángel. Apdo. Postal 47., C.P. 70902., Puerto Ángel, Oaxaca, MÉXICO
**Correo electrónico: pch@angel.umar.mx*

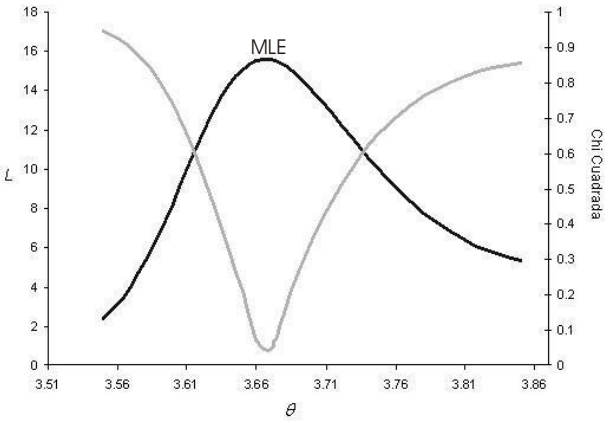

Figura 1. Perfil *L* y el intervalo de confianza para un determinado parámetro θ en un modelo X. Línea oscura (*L*), línea clara (Chi-cuadrada). El valor hipotético MLE para θ es 3.66.

características específicas de *S* en probabilidad clásica y verosimilitud generarán, respectivamente, una estimación puntual de la probabilidad de ocurrencia de θ, y un grupo de probabilidades en las que θ puede ocurrir. Estas últimas son representadas en un perfil de verosimilitud, resaltando en su parte más alta la máxima probabilidad de ocurrencia de θ o estimador de máxima verosimilitud (MLE, por sus siglas en inglés) (Fig. 1).

A diferencia de la verosimilitud, los MC no proporcionan un valor de probabilidad de ocurrencia por parámetro estimado, ya que no consideran una distribución de probabilidad del error (ε), por lo que la detección del mínimo verdadero es más difícil. La verosimilitud considera una distribución de probabilidad para ε, y con base en ésta, se construyen los perfiles de *L* para cada uno de los parámetros estimados, indicando en cada caso el MLE (Fig. 1).

En la figura 1, el perfil *L* indica que θ puede tomar valores entre 3.52 y 3.86; sin embargo, el valor más confiable para θ, que maximiza el ajuste entre Y_i y $\hat{Y}_i$ es el MLE (3.66 con *L* =15.9% de probabilidad). El intervalo de confianza para el MLE (3.62-3.74), se obtiene al interceptar el perfil de verosimilitud con la distribución de Chi-cuadrada según el criterio de Polacheck *et al.* (1993).

Una curva parecida sólo en aspecto a la Chi-cuadrada puede obtenerse para señalar el mínimo en la técnica de MC; sin embargo, en la práctica no es común generarla, en su lugar los parámetros son obtenidos puntualmente, adicionando el grupo de garantías especificadas anteriormente (r, r^2, mapa de residuos y *p*). La tabla I muestra las diferencias más importantes entre las técnicas de MC y *L* en la estimación de parámetros.

Estudio de caso

Para ejemplificar el uso de estas técnicas en la estimación de parámetros, se utilizará el modelo no lineal de crecimiento continuo de von Bertalanffy (1938), como la función problema a la cual se le estimarán los siguientes parámetros: la longitud infinita (L_{inf}), el coeficiente de conversión catabólica (*k*) y la edad (t_o).

Tabla I. Diferencias entre mínimos cuadrados (MC) y verosimilitud (*L*).

MC	*L*
Minimiza el valor del parámetro θ	Maximiza el valor del parámetro θ
No estima la probabilidad del parámetro θ	Estima la probabilidad del parámetro θ
No utiliza la verosimilitud para la estimación de la probabilidad del parámetro θ	Utiliza los mínimos cuadrados y la verosimilitud para la estimación de la probabilidad del parámetro θ
No considera una distribución de probabilidad para el error ε	Considera una distribución de probabilidad para el error ε
No utiliza la desviación estándar del error ε	Usa la desviación estándar del error ε
Considera la suma de cuadrados de los residuos	Considera la suma producto de las probabilidades de los residuos según la distribución de probabilidad asociada a ε

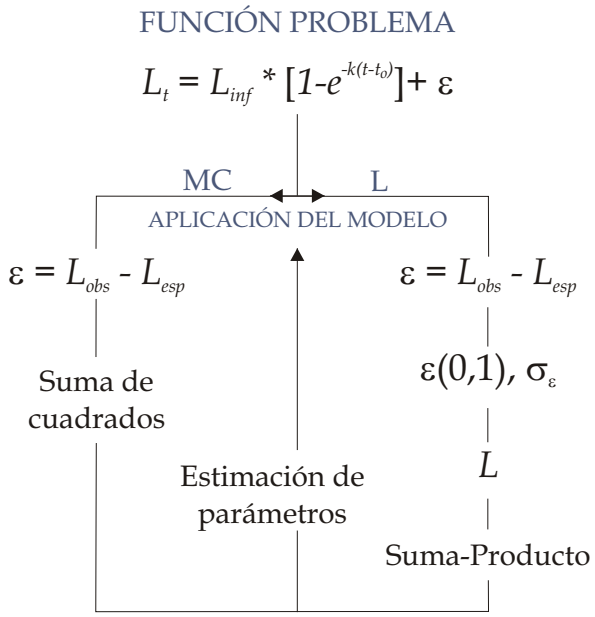

Figura 2. Procedimiento general para desarrollar las técnicas de MC y L en la estimación de los parámetros L_{inf}, k y t_o del modelo no lineal de crecimiento continuo de von Bertalanffy (1938).
ε = error de estimación, ε(0,1) = error de estimación con distribución normal, σ_ε = desviación estándar del error de estimación, Lt_{obs} = longitud observada, Lt_{esp} = longitud esperada.

El desarrollo general para ambas técnicas se muestra en la figura 2, los datos y el procedimiento del cálculo se muestran en las figuras 3 y 5. En el caso de los MC, no se consideró la estimación puntual de los parámetros en cuestión, ya que se pretende comparar los perfiles MC y *L* (Figs. 3c, 5c).

Como puede observarse, para desarrollar la técnica de MC sólo se requiere obtener el residuo entre Lt_{obs} y Lt_{esp}, la estimación de los parámetros se realiza directamente a partir de la suma de cuadrados de los residuos (Figs. 2, 3a-b).

En *L*, los MC están incluidos en ε, los cuales en este caso se consideran con distribución normal, la estimación de los parámetros se realiza directamente con base en la suma producto de *L* (Figs. 2, 5a-b).

La técnica de MC se desarrolló en el programa Excel, en las columnas Edad y Lt_{obs} se localizan los datos de entrada. La columna L_{esp} contiene la función problema que depende

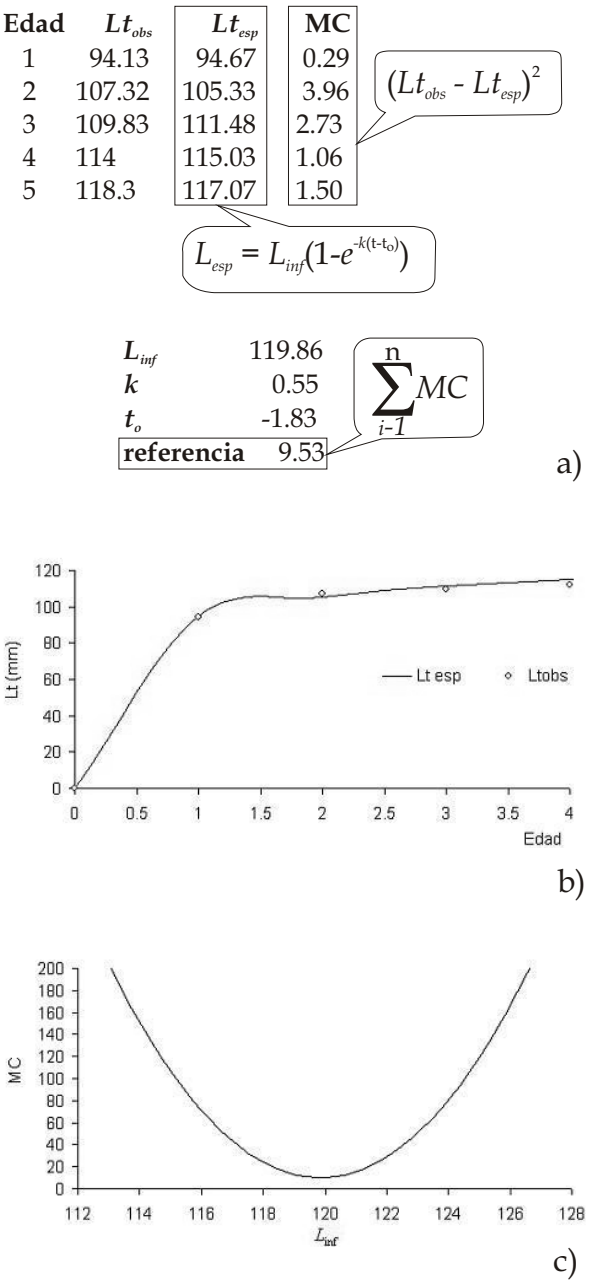

Figura 3. Estimación de los parámetros L_{inf}, k y t_o por MC. a) Tabla de datos y procedimiento del cálculo. b) Ajuste final entre Lt_{esp} y Lt_{obs}. c) Perfil MC para el parámetro L_{inf}.

de un grupo de celdas separadas, en las que se introducen el valor de ensayo para L_{inf}, k y t_o. El cuadrado del residuo se localiza en la columna MC, y la suma de éstos se encuentra en la celda de referencia (Fig. 3a). En el mismo programa Excel, dentro del menú herramientas se selecciona "Solver" (Fig. 4).

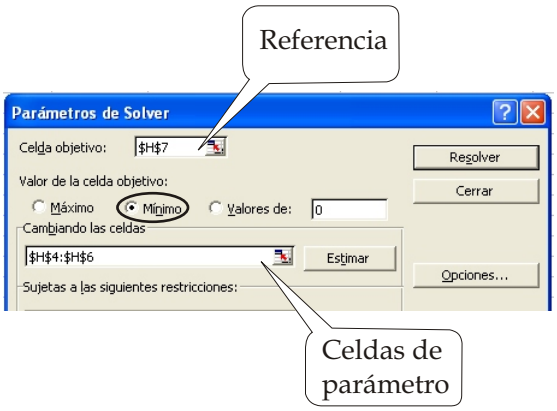

Figura 4. Ventana de la herramienta "Solver" en el programa Excel. Caso MC.

Consideré las indicaciones señaladas en la figura 4 para que "Solver", con base en el mínimo, estime un valor para cada uno de los parámetros considerados. Los resultados y el ajuste final para la función problema se muestran en las figuras 3a y 3b. Para la obtención del perfil MC descrito en la figura 3c y para los demás parámetros consulte a Hilborn & Mangel (1997).

Con respecto a la figura 3b, el ajuste obtenido a través del mínimo entre Lt_{obs} y Lt_{esp} en la función problema, indicó que sólo las longitudes observadas de 0 y 94.13 mm ajustaron correctamente; sin embargo, y aunque próximas a Lt_{esp}, el resto de éstas podrían considerarse con un ligero error de ajuste con respecto al obtenido en la figura 5b.

La técnica *L* se desarrolló en el programa Excel, en las columnas Edad y Lt_{obs} se localizan los datos de entrada. La columna Lt_{esp} contiene la función problema que depende de un grupo de celdas separadas, en las que se introducen el valor de ensayo para L_{inf}, k y t_o. El residuo se localiza en la columna ε, mientras que la función de probabilidad normal está contenida en la columna *L*, la cual considera el valor de la desviación estándar del error de estimación (σ_ε). La suma producto de *L* se localiza en una celda independiente asignada con el mismo nombre, la cual se usó como referencia para "Solver" (Fig. 5a). El nuevo escenario para "Solver" se muestra en la figura 6.

Consideré las indicaciones antes señaladas en la figura 6, para que "Solver", con base en el máximo, estime un valor para cada uno de los parámetros considerados. Los resultados y el ajuste final para la función problema se muestran en las figuras 5a y 5b. Para la obtención del perfil *L* descrito en la figura 5c y para los demás parámetros, consulte a Hilborn & Mangel (1997).

Como puede observarse en la figura 5b, el total de las longitudes observadas ajustaron correctamente con respecto a la línea esperada asignada a la función problema. Por lo que se sugiere, que los parámetros estimados con base en la técnica de *L* son estadísticamente más confiables que los obtenidos por los MC. Las diferencia entre los parámetros estimados por ambas técnicas son, respectivamente: L_{inf} $\pm$2.0 mm, para $k \pm 0.11$ y para $t_o \pm 0.51$.

Es decisión del analista de datos su preferencia por alguna de estas técnicas; sin embargo, en los casos en los que la función problema es más compleja (lineal o no), los MC resultan menos precisos en la estimación de los parámetros involucrados. Por lo anterior los MC sólo son recomendados para desarrollar modelos lineales simples.

Para un mayor entendimiento en el desarrollo de estas técnicas y especialmente para la construcción de los perfiles y los intervalos de confianza de *L*, el lector puede remitirse a: Hilborn & Walters (1992), Punt & Hilborn (1996), Hilborn & Mangel (1997) y Anónimo (2001).

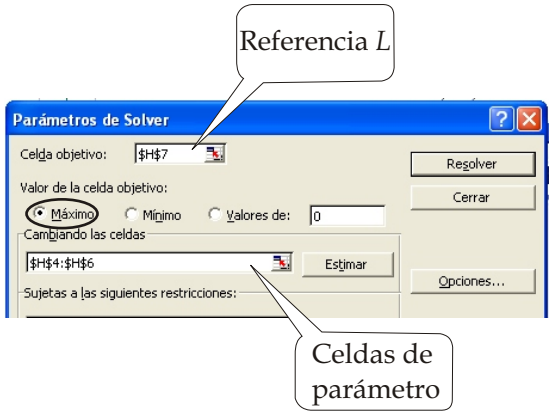

Figura 6. Ventana de la herramienta "Solver" en el programa Excel. Caso *L*.

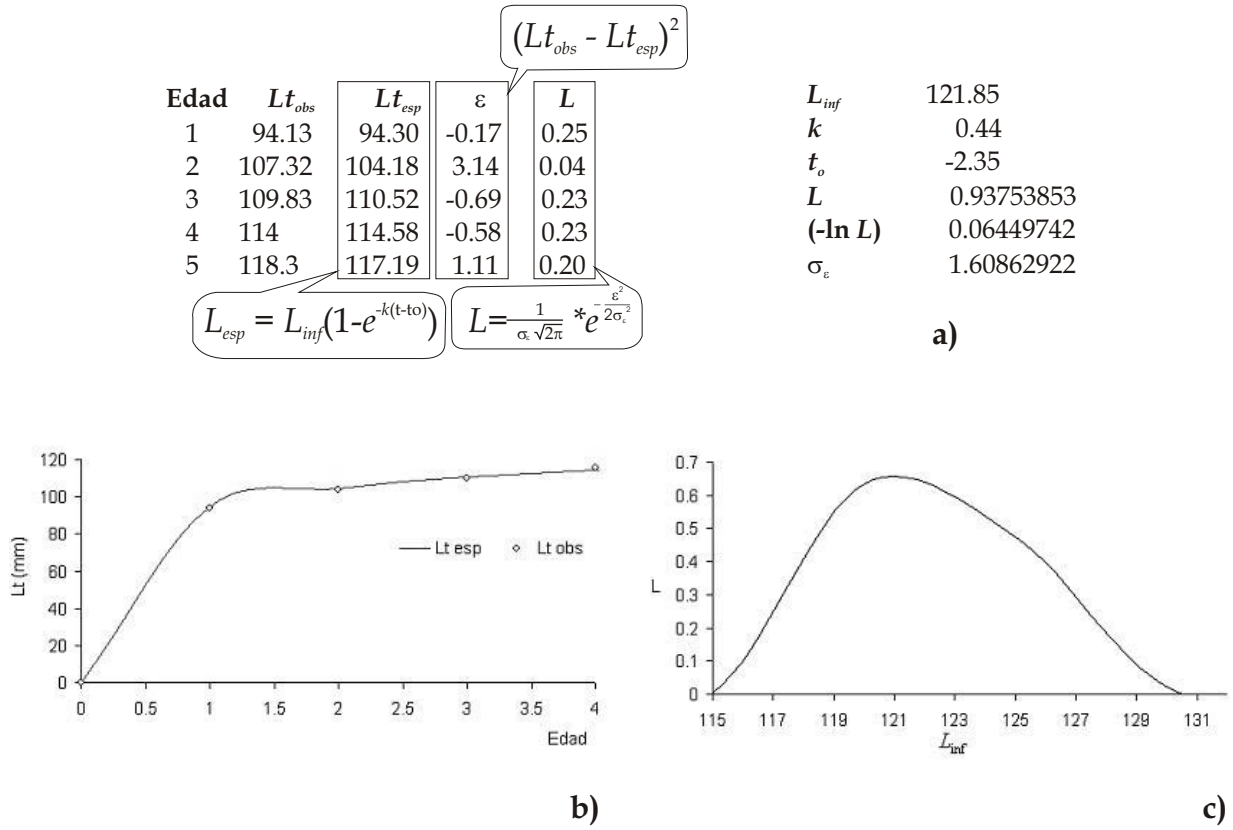

Figura 5. Estimación de los parámetros L_{inf}, k y t_o por L. a) Tabla de datos y procedimiento del cálculo. b) Ajuste final entre Lt_{esp} y Lt_{obs}. c) Perfil L para el parámetro L_{inf}.

Referencias

Anónimo. 2001. Bayes-sa, Bayesian stock assessment methods in fisheries, user's manual. Computerized Information Series. Food and Agriculture Organization of the United Nations (FAO), Biodyn, Roma, 56 pp.

Hilborn, R.R. & M. Mangel. 1997. The ecological detective. Confrontating models with data. Priceton University Press, 330 pp.

Hilborn, R. & C.J. Walters. 1992. Quantitative fisheries stock assessment. Choice, dynamics and uncertainty. Chapman and Hall, New York, 550 pp.

Polacheck, T., R. Hilborn & A. Punt. 1993. Fitting surplus production models: compatating methods and measuring uncertainty. Canadian Journal of Fisheries and Aquatic Sciences 50: 2597-2607.

Punt, A.E. & R. Hilborn. 1996. Biomass dynamic models. Computerized Information Series. Food and Agriculture Organization of the United Nations (FAO), Biodyn, Roma, 62 pp.

Von Bertalanffy L. 1938. A quantitative theory of organic growth. Human Biology 10: 181-213.