¿Qué es el método de estimación de máxima verosimilitud y cómo se interpreta?

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Octubre 2003

www.seh−lelha.org/stat1.htm

Artículo en formato PDF

"En la historia de la estadística existen también piedras angulares y pilares fundamentales, como la distribución de campana de Gauss, el método de mínimos cuadrados, y el principio de máxima verosimilitud". −−

"En la estadística moderna el principio de máxima verosimilitud se ha convertido en una idea sencilla, para algunos incluso evidente. Después de todo, ¿quién puede oponerse a la afirmación de que −entre todas las explicaciones posibles para los datos, se escogerá como la mejor aquella que hace a los datos observados los más probables−? Pero si este principio fuera verdaderamente tan simple, si solo se tratara de eso, ¿por qué muchos de los mayores cerebros de la historia de las matemáticas le han encontrado defectos, desde el siglo dieciocho hasta nuestros días? ¿Cómo es posible que algunos historiadores presenten como primera aparición de esta idea un libro de 1760 de Johann Heinrich Lambert, en el que Lambert solo presenta un sencillo ejemplo y es un ejemplo erróneo?". −−

Stephen M. Stigler, Statistics on the table

Introducción

Muchos procedimientos estadísticos suponen que los datos siguen algún tipo de modelo matemático que se define mediante una ecuación, en la que se desconoce alguno de sus parámetros, siendo éstos calculados o estimados a partir de la información obtenida en un estudio bien diseñado para tal fin. Existen diferentes procedimientos para estimar los coeficientes de un modelo de regresión, o para estimar los parámetros de una distribución de probabilidad. De entre esos procedimientos probablemente el más versátil, ya que se puede aplicar en gran cantidad de situaciones, y por ello uno de los más empleado se conoce con el nombre de "método de máxima verosimilitud" (en inglés "method of maximum likelihood").

Aunque para aquellos que tiene una formación estadística este método es perfectamente conocido y comprendido, sin embargo muchos de los usuarios de los programas estadísticos, que están habituados a calcular modelos de regresión logística, o modelos de supervivencia de riesgo proporcional o de Cox, modelos de Poisson, y muchos otros, desconocen cómo se efectúa la estimación de los coeficientes de esos modelos, por lo que nos parece adecuado dedicar una de éstas páginas mensuales a describir su filosofía e interpretación. Por otro lado, no es infrecuente que empleemos técnicas de forma habitual y mecánica, sin conocer en qué se sustentan y en última instancia en qué consisten realmente: no me cabe ninguna duda que casi todo el mundo tiene claro qué es una distribución de probabilidad normal, pero ¿cuánta gente que utiliza la t de Student sabe qué es realmente eso?

Podemos considerar que el método de máxima verosimilitud, abreviado a menudo como **MLE**, tal y como hoy lo conocemos e interpretamos fue propuesto por Fisher (1890−1962), aunque ya de una forma mucho más artificiosa fue inicialmente atisbado por Bernoulli (1700−1782), cuyo planteamiento fue revisado y

modificado por su coetáneo y amigo el gran matemático Euler (1707−1783). Sin embargo la resolución de los problemas numéricos planteados por este método en la mayor parte de los casos son de tal magnitud que no ha sido posible su amplia utilización hasta la llegada de los modernos ordenadores.

El principio de máxima verosimilitud

Supongamos que se desea estimar la prevalencia en España de personas de más de 50 años con cifras de tensión igual o superior a 160/100 mmHg. Vamos a llamar a esa prevalencia *p* y si se calcula en tanto por 1 será $0 \leq p \leq 1$. Para ello se obtiene una muestra aleatoria y representativa de esa población de tamaño *N*. Supongamos que denotamos con la letra *X* el número de sujetos que presentan cifras tensionales iguales o superiores a los límites fijados en nuestra muestra. El valor concreto que observaremos para *X* puede ser 0 (ningún sujeto), 1, 2, hasta como máximo *N* (todos los sujetos).

En este ejemplo es razonable suponer que la variable aleatoria *X*, número de sujetos con cifras altas de tensión, que observaremos en nuestro estudio (es aleatoria porque si repetimos el trabajo con otra muestra diferente del mismo tamaño es poco probable que el valor observado sea exactamente el mismo) siga una distribución de probabilidad binomial, cuya fórmula es la siguiente:

$$P(X) = C_{X,N} \cdot p^{X} \cdot (1-p)^{N-X} \qquad [1]$$

donde $C_{X,N}$ es el número combinatorio que se calcula como $N!/X!(N-X)!$

Para simplificar la exposición, supongamos que se utiliza una muestra de $N=200$ sujetos. Una vez que efectuamos el estudio conocemos el valor de *X* y podemos calcular la probabilidad de observar exactamente ese valor para diferentes prevalencias posibles en la población. Esa probabilidad que hemos llamado *P(X)* es función de *N*, *X* y *p*; luego conocidas las dos primeras variables podemos probar con distintos valores de prevalencia *p* y determinar qué valor de prevalencia en la población nos conduce a una mayor *P(X)*, o lo que es lo mismo para qué valor real de la prevalencia en la población es más probable que observemos ese valor concreto de *X* en una muestra aleatoria de *N* sujetos.

Supongamos que el número *X* de pacientes con cifras de tensión iguales o superiores al límite prefijado es de *60*. Podemos plantear cuál es la probabilidad de obtener *60* sujetos hipertensos en una muestra de *200* personas si la prevalencia real fuera de $p=0.2$. Substituyendo esos valores en la ecuación [1] obtenemos $P(X)=0.00022$. Si la prevalencia real fuera $p=0.3$ el valor de *P(X)* calculado sería *0.06146*, mayor que el anterior; y si $p=0.4$ entonces $P(X)=0.00082$, que también es menor que el calculado para $p=0.3$. El método de máxima verosimilitud nos dice que escogeremos como valor estimado del parámetro aquél que tiene mayor probabilidad de ocurrir según lo que hemos observado, es decir aquél que es más compatible con los datos observados, siempre suponiendo que es correcto el modelo matemático postulado.

Obviamente no se trata de ir probando diferentes valores del parámetro, en este caso de la prevalencia *p*, para ver cuál es que proporciona un mayor valor de verosimilitud. La ecuación [1], una vez fijados en nuestro estudio *N* y *X*, es únicamente función de *p*, por lo que podemos representar en una gráfica el resultado de sustituir diferentes valores de *p* en esa ecuación y obtendremos una gráfica como la de la figura 1, donde se ve que esa función tiene su valor máximo para *0.3*. Luego con ese modelo matemático *0.3* es el valor de la prevalencia más verosímil de acuerdo con los datos que hemos obtenido en nuestro estudio *(N=200, X=60)*

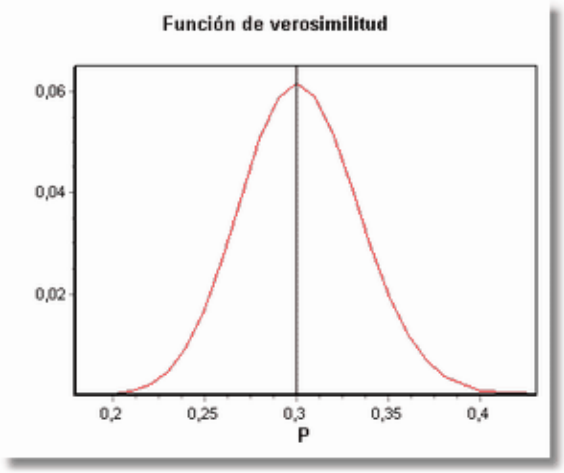

Figura 1

Desde luego para poder representar esa gráfica, salvo que dispongamos de un programa adecuado, también tenemos que calcular cada pareja de valores (Verosimilitud,p) y además después dibujarlos. Sin embargo existe un procedimiento matemático para determinar el punto máximo o mínimo de una ecuación, que consiste en calcular la derivada de la función e igualar a cero. Se trata en realidad de determinar, de forma matemática, la pendiente en cada punto (eso es la derivada) y en el punto máximo sabemos que la pendiente es cero (basta mirar la figura 1). Si el lector todavía recuerda como se hacía eso y prueba con la ecuación [1], puede comprobar que el máximo de esa función se obtiene para el valor de *p* calculado como *X/N*, que no es mas que la proporción de sujetos hipertensos observada en nuestro estudio, lo que por otro lado parece obvio, pero resulta bastante tranquilizador que las matemáticas corroboren algo que nos parece obvio, a saber que la estimación más verosímil de una proporción a partir de una muestra aleatoria corresponde al cociente entre el número de sucesos partido por el tamaño de la muestra. Sin embargo este razonamiento es general y hay muchos casos en el que el resultado no es tan sencillo y sí es imprescindible la matemática para estimar los parámetros.

Estimación de modelos por el método de máxima verosimilitud

El método de máxima verosimilitud se utiliza por ejemplo para estimar los coeficientes de un modelo logístico de regresión, en el que se calcula la probabilidad de que ocurra un determinado suceso mediante la siguiente ecuación:

$$p = \frac{1}{1+\exp[-(b_0 + b_1 x_1 + \ldots + b_k x_k)}$$ [2]

donde *p* es la probabilidad de que ocurra el suceso de interés y x_i son los posibles factores (factores de riesgo) que se piensa que están relacionados con la probabilidad de que el suceso se produzca.

Ahora a partir de los datos de la muestra, para los que hemos observado si se ha producido o no el suceso, y a partir de los valores de los factores de riesgo en cada caso de la muestra, se trata de estimar los valores de los coeficientes b_i en el modelo para cada factor de riesgo, lo que entre otras cosas nos permite calibrar el efecto de esos factores en la probabilidad de que el suceso ocurra. Si denominamos de forma compacta a esos coeficientes con la letra *b* (vector de valores), y dado que los valores de los factores *x* son conocidos para cada sujeto, la probabilidad *p* es función de los coeficientes *b*, y lo representamos como $p=f(b)$.

Si *p* es la probabilidad de que ocurra el suceso, la de que NO ocurra será $1-p$, y entonces en los sujetos en los que ocurrió el suceso vendrá dada por $p(x_i)$, mientras que para un sujeto en el que NO ocurre el suceso, se calcula como $1-p(x_i)$. Siendo ambas expresiones función de *b*.

Si la muestra es aleatoria y las observaciones son independientes entre sí, la probabilidad de que un sujeto de la muestra experimente el suceso es independiente de lo que le ocurra a cualquier otro, por lo que la

probabilidad conjunta se calcula como el producto de las probabilidades individuales y de esa forma obtenemos la función de verosimilitud, que tiene en cuenta todos los datos de forma global, y será función únicamente de los coeficientes. De igual manera que antes se calculará la derivada de esa función, se iguala a cero y se obtienen los valores de los coeficientes que maximizan esa función. Aunque esto que se dice fácil, al menos en el modelo logístico, es algo más complicado de efectuar que de narrar. Pero de eso hablaremos en otra ocasión.

Interpretación de los resultados en el método de máxima verosimilitud

Al combinar observaciones independientes, hemos visto que en el cálculo de la función de verosimilitud interviene el producto de las probabilidades individuales, por lo que habitualmente interesa tomar logaritmos, ya que éstos transforman los productos en sumas y los cocientes en restas. Así habitualmente veremos en las salidas de los programas de ordenador el término **Log−likehood**, que no es más que el **logaritmo de la verosimilitud**. Al tratarse de productos de probabilidades la función de verosimilitud será siempre menor que 1 y por tanto su logaritmo será negativo.

La función de verosimilitud nos permite comparar modelos, por ejemplo dos modelos en el que en uno de ellos se incluye una variable adicional con respecto al primer modelo. Las diferencias en la función de verosimilitud se alteran arbitrariamente con la escala de medida, por lo que la forma adecuada de compararlas es mediante cocientes. De ahí que cuando se comparan modelos que han sido estimados mediante este procedimiento se hable de **cociente de verosimilitud** (**likelihood ratio**).

Cuando se trata de la estimación de modelos resulta de utilidad el concepto de **modelo saturado**. Un modelo se denomina saturado cuando utiliza tantos parámetros como observaciones hemos efectuado y por tanto se ajusta perfectamente a los datos. Podemos comparar el modelo actualmente estimado con ese modelo teórico perfecto mediante la expresión:

$$D = -2\ln\left(\frac{Verosimilitud\ \ m.actual}{Verosimilitud\ \ m.saturado}\right)$$

esa cantidad se denomina **desviación** (**deviance** en inglés; en algún lugar la he visto traducida cono "*desvianza*", término que no creo que exista en nuestro idioma y que a mi particularmente no me suena bien).

La desviación nos permite comparar modelos, por ejemplo un modelo que incluye una variable adicional:

$$G=D(\text{modelo 1 sin la variable}) - D(\text{modelo 2 con la variable}) = -2\ln\left(\frac{Verosimilitud\ \ 1}{Verosimilitud\ \ 2}\right)$$

que se distribuye según una chi^2 con grados de libertad igual a la diferencia de parámetros entre modelos, que este caso es 1 grado de libertad. Se le denomina **contraste de verosimilitud**. Si el contraste resulta ser no significativo aceptamos que la incorporación de la nueva variable no mejora sensiblemente la verosimilitud del modelo y por tanto no merece la pena incluirla en él.

También en las salidas de los programas suele aparecer el término likelihood ratio o cociente de verosimilitud para un modelo, sin que se especifique que se esté contrastando con otro diferente. En estos casos el contraste es frente al modelo que sólo incluye el término constante y por tanto no se consideran las variables *X* o los factores de riesgo, y se compara con el modelo que sí incluye las variables, por lo que ahora esa cantidad se distribuye según una chi^2 con grados de libertad igual al número de variables incluidas en el modelo, que es la diferencia frente al modelo con solo la constante. Al igual que antes, si el contraste resulta no significativo pensamos que incluir el conocimiento de las variables X no mejora significativamente la verosimilitud del modelo y por lo tanto se trata de un modelo sin utilidad.

Añadiendo más términos, más variables, a un modelo la función de verosimilitud mejorará y si la muestra es grande será difícil distinguir mediante el contraste del cociente de verosimilitud entre una mejora "real" y una aportación trivial. El modelo perfecto no existe, puesto que todos constituyen simplificaciones de la realidad y siempre son preferibles modelos con menos variables, puesto que además de ser más sencillos, son más estables y menos sometidos a sesgo. Por ello se han propuesto otras medidas de contraste entre modelos que penalizan en alguna medida que éstos tengan muchos parámetros.

Las más conocidas y que suelen figurar en las salidas de ordenador son el **criterio de información de Akaike**, **AIC**, y **criterio de información bayesiano**, **BIC**.

$$AIC = -2(\ln \text{verosimilitud} - n^{o}\ \text{parámetros})$$

En principio el criterio de selección será escoger modelos con valores más bajos de AIC.

La fórmula para el BIC es similar, así como su interpretación:

$$BIC = G - gl \cdot \ln N$$

donde G es el cociente de verosimilitud, gl son los grados de libertad y N el tamaño de la muestra. También escogeremos modelos con menor valor de BIC.

Epílogo

Tal y como hemos planteado el método de máxima verosimilitud es un procedimiento que permite estimar los parámetros de un modelo probabilístico, o los coeficientes de un modelo matemático, de tal manera que sean los más probables a partir de los datos obtenidos. También hemos visto que por ello nos permite comparar diferentes modelos, incluyendo o no variables en el mismo.

Hay que tener bien claro que en el método además de intervenir la información aportada por los datos, se está postulando un modelo matemático para éstos, como puede ser por ejemplo el modelo logístico o un modelo de supervivencia, y que los parámetros estimados se calculan considerando la información aportada por los datos de acuerdo a ese modelo. Si el modelo propuesto no fuera adecuado el método tampoco lo será. Quiere esto decir que la razón de verosimilitud no nos proporciona información suficiente en cuanto a la bondad de ajuste, que habrá que verificar convenientemente por otros métodos.

No he logrado encontrar enlaces con una exposición sencilla del método de máxima verosimilitud. A continuación indico el enlace que me ha parecido de más fácil comprensión.

Enlaces

Maximum Likelihood Estimation
S. Purcell

Indice de artículos

Principio de la página