Insumos de estimación

Algoritmos

Nicolás Rivera Garzón

Departamento de Estadística
Universidad Nacional de Colombia

Abril, 2021

Contenido

Problemas Algorítmicos: MLE

- Las estimaciones de máxima verosimilitud pueden no estar dadas explícitamente por fórmulas, sino implícitamente como soluciones de sistemas de ecuaciones no lineales.
- En el modelo de regresión clásico con matriz de diseño de rango completo, la fórmula del parámetro es fácil de escribir simbólicamente pero no es fácil de evaluar si el rango es grande.

Modelo de regresión lineal simple (múltiple)

Para los datos $(z_i, Y_i; i = 1, \ldots, n)$, el modelo en forma matricial es:

$$Y = Z_D\beta + \epsilon$$

donde Z_D es la matriz de diseño. La parametrización es identificable si y solo si Z_D es de rango d o de manera equivalente, si $Z_D^t Z_D$ es de rango completo d. En ese caso la solución es:

$$\hat{\beta} = [Z_D^t Z_D]^{-1} Z_D^t Y$$

Método de Bisección

Si la verosimilitud es diferenciable y la ubicación del máximo global está en el interior del espacio de parámetros, encontrar la solución se reduce a encontrar el resultado de las ecuaciones normales de verosimilitud. Los pasos del método son

1. Encontrar dos puntos, uno en el que la derivada está por debajo de cero y otro que está por encima.
2. Tomar la mitad del intervalo formado por estos puntos y cortarlo a la mitad.
3. Evaluar el valor de la derivada en el medio y reemplazar por el punto medio uno de los puntos originales que tiene el mismo signo del valor de la derivada.
4. Continuar hasta que la longitud del intervalo tenga la precisión deseada o se alcance el cero antes de eso.

Bisección: Pseudocódigo

Dada una función continua f en (a, b), f estrictamente creciente y $f(a+) < 0 < f(-b)$, entonces el **teorema del valor intermedio** garantiza que existe una única $x^* \in (a, b)$ tal que $f(x^*) = 0$. Con esto en mente, el método de bisección encuentra x^* con el pseudocódigo:

1. Encontrar $x_0 < x_1, f(x_0) < 0 < f(x_1)$, tomando $|x_0|, |x_1|$ suficientemente grandes, iniciar el algoritmo con $x^+_{old} = x_1, x^-_{old} = x_0$
2. Si $|x^+_{old} - x^-_{old}| < 2\epsilon$, entonces $x_{final} = \frac{1}{2}(x^+_{old} + x^-_{old})$ y declarar x_{final}
3. De otro modo, $x_{new} = \frac{1}{2}(x^+_{old} + x^-_{old})$
4. Si $f(x_{new}) = 0$, entonces, $x_{final} = x_{new}$ y declarar x_{final}
5. Si $f(x_{new}) < 0$, entonces, $x^-_{old} = x_{new}$
6. Si $f(x_{new}) > 0$, entonces, $x^+_{old} = x_{new}$
7. Volver al paso 2

El algoritmo de bisección para en la solución x_{final} tal que:

$$|x_{final} - x^*| \leq \epsilon$$

Bisección para la familia exponencial uniparamétrica

Teorema 2.4.1

Sea $p(x, \eta)$ una familia exponencial canónica de un parámetro generada por (T, H), entonces, el estimador $\hat{\eta}$ de máxima verosimilitud se puede encontrar (con tolerancia ϵ) por el método de bisección aplicado a:

$$f(\eta) \equiv E_\eta T(X) - t_0$$

- Para la familia gamma de parámetro de forma con $X_1, \ldots, X_n$ i.i.d. $\Gamma(\theta, 1)$ de la forma:

$$p(x; \theta) = \Gamma^{-}1(\theta) x^{\theta - 1} e^{-x}, x > 0, \theta > 0$$

Ya que $T(X) = \sum_{i=0}^{n} log(X_i)$ tiene una densidad para todo n tal que el estimador de máxima verosimilitud existe. De esta forma, se tiene que solucionar la ecuación:

$$\frac{\Gamma^{'}(\theta)}{\Gamma(\theta)} = \frac{T(X)}{n}$$

que se puede evaluar por el método de bisección por el teorema anterior.

Algoritmo de ascenso por coordenadas: extensión a parámetros multidimensionales

El problema a considerar es resolver numéricamente, para la familia exponencial canónica de k parámetros, la ecuación normal de:

$$E_\eta(T(X)) = \dot{A} = t_0$$

donde el estimador de máxima verosimilitud de $\hat{\eta} \equiv \hat{\eta}(t_0)$ existe. Aunque el algoritmo sea lento, siempre converge a $\hat{\eta}$.

El método de ascenso por coordenadas es el algoritmo de bisección aplicado a cada coordenada en forma iterativa de reciclaje.

Ascenso por coordenadas: Pseudocódigo

Para encontrar $\hat{\eta}$ se siguen los siguientes pasos:

1. Comenzar con un $\hat{\eta}^0 = (\hat{\eta}_1^0, \ldots, \hat{\eta}_k^0)$ arbitrario, preferiblemente con una estimación ad hoc.
2. Resolver las ecuaciones para las primeras derivadas de la forma:

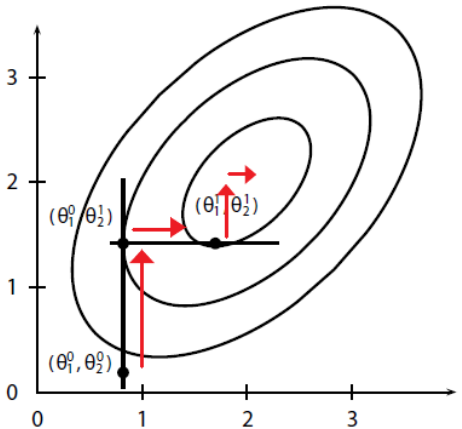
$$\hat{\eta}_1^1 : \frac{\partial}{\partial \eta_1} A(\eta_1, \hat{\eta}_2^0, \ldots, \hat{\eta}_k^0) = t_1$$
$$\hat{\eta}_2^1 : \frac{\partial}{\partial \eta_2} A(\hat{\eta}_1^1, \eta_2, \hat{\eta}_3^0, \ldots, \hat{\eta}_k^0) = t_2$$
$$\hat{\eta}_k^1 : \frac{\partial}{\partial \eta_k} A(\hat{\eta}_k^1, \hat{\eta}_2^1, \ldots, \hat{\eta}_k^0) = t_k$$

3. Definir $\hat{\eta}^{01} \equiv (\hat{\eta}_1^1, \hat{\eta}_2^0, \ldots, \hat{\eta}_k^0), \hat{\eta}^{02} \equiv (\hat{\eta}_1^1, \hat{\eta}_2^1, \hat{\eta}_3^0, \ldots, \hat{\eta}_k^0)$ y sucesivamente.
4. Repetir el proceso hasta eventualmente encontrar $\hat{\eta}^{(r)}$ con una tolerancia igual a:

$$|\hat{\eta}^{jl} - \hat{\eta}^{j(l-1)}| \leq \epsilon$$

¿Por qué funciona?

- La log-verosimilitud es cóncava.
- El algoritmo encuentra el máximo a lo largo de una de las coordenadas, mientras que todas las demás se fijan de modo que el valor de la verosimilitud aumenta en cada paso del algoritmo.
- La log-verosimilitud está limitada en el interior del conjunto de parámetros, por lo que tiene que haber un límite.
- Gracias a ello, $\hat{\eta}^{(r)}$ debe estar en una bola cerrada en el interior del espacio de parámetros.



Algoritmo Newton-Raphson

- Este método requiere el cálculo de la inversa de la hessiana, lo que puede contrarrestar su ventaja en la velocidad de convergencia sobre el ascenso por coordenadas, cuando converge.
- La lógica de su pseudocódigo es sencilla, si $\hat{\eta}_{old}$ es el valor actual del algoritmo entonces:

$$\hat{\eta}_{new} = \hat{\eta}_{old} - \ddot{A}^{-1}(\dot{A}(\hat{\eta}_{old}) - t_0)$$

Si $\hat{\eta}_{old}$ está cerca de la raíz $\hat{\eta}$ de $\dot{A}(\hat{\eta}) = t_0$, entonces por la expansión de $\dot{A}(\hat{\eta})$ alrededor de $\hat{\eta}_{old}$ se obtiene:

$$t_0 - \dot{A}(\hat{\eta}_{old}) = \dot{A}(\hat{\eta}) - \dot{A}(\hat{\eta}_{old}) \simeq \ddot{A}(\hat{\eta}_{old})(\hat{\eta} - \hat{\eta}_{old})$$

$\hat{\eta}_{new}$ es la solución para la aproximación dada por la ecuación. El algoritmo tiene la propiedad que para n grande, $\hat{\eta}_{new}$ se comporta aproximadamente como MLE después de un paso.

Newton-Raphson para MLE y familia exponencial

- Para una función de log-verosimilitud $l(\theta)$:

$$\hat{\theta}_{new} = \hat{\theta}_{old} - \ddot{l}_{-1}(\hat{\theta}_{old})\dot{l}(\hat{\theta}_{old})$$

- Para una familia exponencial canónica:

$$\hat{\eta}_{new} = \hat{\eta}_{old} - \ddot{A}^{-1}(\hat{\eta}_{old})(\dot{A}(\hat{\eta}_{old}) - t_0)$$

Cuando las probabilidades no son cóncavas, se siguen empleando métodos como bisección, ascenso coordinado y Newton-Raphson. Existe una clara posibilidad de no convergencia o convergencia a un máximo local en lugar de global.

Algoritmo esperanza-maximización (EM): Motivación

Existen muchos modelos que tienen la siguiente estructura:

- Hay observaciones ideales, $X \sim P(\theta)$ con densidad $p(x;\theta), \theta \in \Theta \subset R^d$.
- Su función de log-verosimilitud $l_{p,x}(\theta)$ es fácil de maximizar.
- Sin embargo, se observan datos incompletos dados por $S \equiv S(X) \sim Q_\theta$ con densidad $q(s,\theta)$ donde $l_{q,s}(\theta) = log(q(s,\theta))$ es difícil de maximizar. Su función no es cóncava y es compleja de computar,
- Una forma de pensar en tales problemas es en términos de S como datos incompletos y que representan una parte de X, y su "reconstrucción" es parte del proceso de estimación por máxima verosimilitud de θ.

Algoritmo EM

1. Iniciar con $\theta_{old} = \theta_0$
2. Sea:
$$J(\theta|\theta_0) \equiv E_{\theta_0}(\log \frac{p(X,\theta)}{p(X,\theta_0}|S(X) = S)$$
3. Paso E: evaluar $J(\theta|\theta_0)$ para tantos valores de θ como sea necesario.
4. Paso M: Maximizar $J(\theta|\theta_0)$ como una función de θ.
5. Definir $\theta_{new} = \arg\max J(\theta|\theta_0)$ y reiniciar $\theta_{old} = \theta_{new}$ y repetir el proceso.

Lema 2.4.1

Si θ_{new},θ_{old} se definen como en el pseudocódigo y $S(X) = s$ entonces:

$$q(s, \theta_{new}) \geq q(s, \theta_{new})$$

La igualdad se cumple si la la distribución condicional de X dada $S(X) = s$ es la misma para θ_{new} como θ_{old} y θ_{old} maximiza $J(\theta|\theta_{old})$

EM para familias exponenciales

Teorema 2.4.3

Sea $P_\theta : \theta \in \Theta$ una familia exponencial canónica generada por (T, h) y sea $S(X)$ cualquier estadístico, entonces:

1. El algoritmo EM consiste de la alternancia de:

$$\dot{A}(\theta_{new}) = E_{\theta_{old}}(T(X)|S(X) = s)$$

$$\theta_{old} = \theta_{new}$$

2. Si:

$$\dot{A}(\theta) = E_{\theta}(T(X)|S(X) = s)$$

tiene una única solución que converge a $\hat{\theta}^*$ que es necesariamente un máximo local de $q(s, \theta)$.

Referencias

- Bickel, P., Doksum, K. (2015). Mathematical statistics. Boca Raton: CRC Press.
- El repositorio con la presentación y las simulaciones se encuentra en GitHub.