2. Cómputo Bayesiano

- Recordemos que $f(\theta|\underline{x})$, la distribución final de θ, contiene toda la información relevante sobre la incertidumbre asociada a los parámetros de un modelo.

- La toma de decisiones (inferencia) consiste en resumir u obtener una característica numérica de la distribución final $f(\theta|\underline{x})$, como momentos o distribuciones marginales. Sin embargo, en algunos casos, la forma de $f(\theta|\underline{x})$ es tan complicada que no es posible integrarla de manera analítica, por lo que es necesario recurrir a aproximaciones ya sean analíticas o numéricas.

2.1 Aproximaciones analíticas e integración numérica

- APROXIMACIONES ANALÍTICAS: Las aproximaciones analíticas que veremos están basadas en argumentos asintóticos. Las más usadas son: aproximación normal asintótica y método de Laplace.

1) *Aproximación normal asintótica*: Por simplicidad, supondremos que θ es un parámetro de dimensión 1, el caso general es análogo.

   - Teorema: Aproximación normal asintótica.

     Bajo ciertas condiciones de regularidad y para valores de n suficientemente grandes, la distribución final de θ *es aproximadamente normal con media* $\hat{\theta}$ *y varianza* $V(\hat{\theta})$*, es decir,*

$$f(\theta|\underline{x}) \approx N(\theta|\hat{\theta}, V(\hat{\theta}))$$

donde, $\hat{\theta}$ *es la moda de* $f(\theta|\underline{x})$ *y*

$$V(\hat{\theta}) = -\left\{\frac{\partial^2 \log f(\hat{\theta}|\underline{x})}{\partial \theta^2}\right\}^{-1}.$$

DEM.

Sea $f_x(\theta) = f(\underline{x}|\theta) f(\theta)$ el "kernel" o núcleo de la distribución final de θ

Desarrollando $\log f_x(\theta)$ en serie de Taylor alrededor de $\hat{\theta}$ tenemos

$$\log f_x(\theta) = \log f_x(\hat{\theta}) + \frac{\partial \log f_x(\hat{\theta})}{\partial \theta}(\theta - \hat{\theta}) + \frac{1}{2}\frac{\partial^2 \log f_x(\hat{\theta})}{\partial \theta^2}(\theta - \hat{\theta})^2 + \cdots$$

como $\hat{\theta}$ es también la moda de $f_x(\theta)$ y usando la expresión para $V(\hat{\theta})$,

$$\log f_x(\theta) \approx \log f_x(\hat{\theta}) - \frac{1}{2V(\hat{\theta})}(\theta - \hat{\theta})^2$$

$$f_x(\theta) \approx f_x(\hat{\theta}) \exp\left\{-\frac{1}{2V(\hat{\theta})}(\theta - \hat{\theta})^2\right\}.$$

Recordemos que $f(\theta|\underline{x}) = \frac{f_x(\theta)}{\int f_x(\theta) d\theta}$, entonces

$$\int f_x(\theta) d\theta \approx \int f_x(\hat{\theta}) \exp\left\{-\frac{1}{2V(\hat{\theta})}(\theta - \hat{\theta})^2\right\} d\theta = f_x(\hat{\theta})(2\pi)^{1/2}\{V(\hat{\theta})\}^{1/2}.$$

Finalmente,

$$f(\theta|\underline{x}) \approx (2\pi)^{-1/2}\{V(\hat{\theta})\}^{-1/2} \exp\left\{-\frac{1}{2V(\hat{\theta})}(\theta - \hat{\theta})^2\right\}.$$

- EJEMPLO 12: Sea $X_1, X_2, ..., X_n$ una m.a. de Ber(θ). Sea θ~Beta(a,b) la distribución inicial de θ. Sabemos que $\theta|\underline{x} \sim \text{Beta}(a_1, b_1)$, donde

$$a_1 = \alpha + \sum_{i=1}^{n} X_i \quad y \quad b_1 = \beta + n - \sum_{i=1}^{n} X_i \, .$$

Para obtener la aproximación normal asintótica de $f(\theta|\underline{x})$ tenemos,

$$f(\theta|\underline{x}) = c\theta^{a_1-1}(1-\theta)^{b_1-1} I_{(0,1)}(\theta)$$

$$\log f(\theta|\underline{x}) = \log c + (a_1 - 1)\log\theta + (b_1 - 1)\log(1-\theta) + \log I_{(0,1)}(\theta)$$

$$\frac{\partial}{\partial\theta}\log f(\theta|\underline{x}) = \frac{a_1 - 1}{\theta} - \frac{b_1 - 1}{1-\theta}$$

igualando a cero y despejando θ tenemos que $\hat{\theta} = \dfrac{a_1 - 1}{a_1 + b_1 - 2}$.

$$\frac{\partial^2}{\partial\theta^2}\log f(\theta|\underline{x}) = -\frac{a_1 - 1}{\theta^2} - \frac{b_1 - 1}{(1-\theta)^2} \Rightarrow V(\theta) = \left\{ \frac{a_1 - 1}{\theta^2} + \frac{b_1 - 1}{(1-\theta)^2} \right\}^{-1}$$

$$\therefore \; V(\hat{\theta}) = \frac{(a_1 - 1)(b_1 - 1)}{(a_1 + b_1 - 2)^3}$$

Por lo tanto, $f(\theta|\underline{x}) \approx N(\theta|\hat{\theta}, V(\hat{\theta}))$.

2) *Aproximación de Laplace*: Supongamos que θ es de dimensión 1. La aproximación de Laplace permite aproximar integrales de la forma

$$I = \int q(\theta)\exp\{-nh(\theta)\}d\theta$$

donde $q(\cdot)$ y $h(\cdot)$ son dos funciones suaves de θ.

- Teorema: Aproximación de Laplace.

  Sea I *una integral de la forma antes mencionada. Suponga que la función* $h(\cdot)$ *tiene un mínimo en* $\hat{\theta}$. *Para valores suficientemente grandes de n, la integral* I *puede ser aproximada por*

$$\hat{I} = q(\hat{\theta})(2\pi/n)^{1/2}\left|\Sigma(\hat{\theta})\right|^{1/2}\exp\{-nh(\hat{\theta})\},$$

donde,

$$\Sigma(\hat{\theta}) = \left\{\frac{\partial^2 h(\theta)}{\partial\theta^2}\right\}^{-1}.$$

DEM.

La aproximación de Laplace se basa en la expansión en serie de Taylor tanto de h(·) como de q(·) alrededor de $\hat{\theta}$.

Desarrollando h(θ) alrededor de $\hat{\theta}$ tenemos,

$$h(\theta) = h(\hat{\theta}) + \frac{\partial h(\hat{\theta})}{\partial\theta}(\theta-\hat{\theta}) + \frac{1}{2}\frac{\partial^2 h(\hat{\theta})}{\partial\theta^2}(\theta-\hat{\theta})^2 + \cdots,$$

despreciando los términos de orden mayor a 2 y usando $\Sigma(\hat{\theta})$ tenemos

$$\exp\{-nh(\theta)\} \approx \exp\{-nh(\hat{\theta})\}\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta-\hat{\theta})^2\right\}.$$

De manera similar, al desarrollar q(θ) alrededor de $\hat{\theta}$ tenemos,

$$q(\theta) = q(\hat{\theta}) + \frac{\partial q(\hat{\theta})}{\partial\theta}(\theta-\hat{\theta}) + \cdots.$$

Entonces, el integrando de I puede escribirse como

$$q(\theta)\exp\{-nh(\theta)\} = \left\{q(\hat{\theta}) + \frac{\partial q(\hat{\theta})}{\partial\theta}(\theta-\hat{\theta})\right\}\exp\{-nh(\hat{\theta})\}\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta-\hat{\theta})^2\right\}$$

Finalmente, notemos que

$$\int\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta-\hat{\theta})^2\right\}d\theta = (2\pi/n)^{1/2}\left|\Sigma(\hat{\theta})\right|^{1/2},$$

$$\int(\theta-\hat{\theta})\exp\left\{-\frac{n}{2\Sigma(\hat{\theta})}(\theta-\hat{\theta})^2\right\}d\theta = 0$$

de manera que $I \approx q(\hat{\theta})\exp\{-nh(\hat{\theta})\}(2\pi/n)^{1/2}\left|\Sigma(\hat{\theta})\right|^{1/2}$.

- Notemos que en general, una integral dada

$$\int f(\theta)d\theta$$

puede escribirse como

$$\int q(\theta)\exp\{-nh(\theta)\}d\theta$$

para distintas funciones q(·) y h(·). Para un valor fijo de n, la aproximación de Laplace depende tanto de la elección particular de estas funciones como de la parametrización que se utilice.

- EJEMPLO 13: Supongamos que se desea calcular

$$E\{g(\theta)|\underline{x}\} = \int g(\theta)f(\theta|\underline{x})d\theta\,.$$

Como $f(\theta|\underline{x}) \propto f_x(\theta) = f(\underline{x}|\theta)f(\theta)$ entonces,

$$E\{g(\theta)|\underline{x}\} = \frac{\int g(\theta)f_x(\theta)d\theta}{\int f_x(\theta)d\theta}\,.$$

Sean $q(\theta) = g(\theta)$ y $\exp\{-nh(\theta)\} = f_x(\theta) \Rightarrow h(\theta) = -\frac{1}{n}\log f_x(\theta)$.

Aproximando el numerador y el denominador por separado tenemos

$$\int g(\theta)f_x(\theta)d\theta \approx g(\hat{\theta})(2\pi/n)^{1/2}|\Sigma(\hat{\theta})|^{1/2}f_x(\hat{\theta}),\ \text{y}$$

$$\int f_x(\theta)d\theta \approx (2\pi/n)^{1/2}|\Sigma(\hat{\theta})|^{1/2}f_x(\hat{\theta}),$$

donde $\hat{\theta}$ es el mínimo de h(θ). Por lo tanto,

$$\hat{E}\{g(\theta)|\underline{x}\} = g(\hat{\theta}).$$

- INTEGRACIÓN NUMÉRICA: Los métodos de integración numérica, también conocidos como métodos de cuadratura, permiten calcular eficientemente algunas características de la distribución final de θ cuando la dimensión de éste es pequeña. Supongamos que θ es de dimensión 1.

o Definición: Regla de integración.

Sea $f(\cdot)$ *una función suave y supongamos que se desea calcular la integral*

$$I = \int_a^b f(\theta) d\theta .$$

Una regla de integración numérica está definida por un conjunto de nodos, $\{\eta_i\}_{i=1}^N$ *y un conjunto asociado de pesos o ponderaciones,* $\{u_i\}_{i=1}^N$ *tales que*

$$I \approx \sum_{i=1}^N u_i f(\eta_i).$$

- Sean $a = \theta_0 < \theta_1 < \cdots \theta_N = b$ los valores de N+1 puntos distribuidos en el intervalo [a,b]. En particular, si los puntos son equidistantes, entonces

$$\theta_i = \theta_0 + ih \text{ , (i=1,...,N)}$$

donde $h = (b-a)/N$. En otras palabras,

$$\theta_i = a + \frac{i(b-a)}{N} \text{ , (i=1,...,N).}$$

1) *Regla del punto medio*. Está dada por

$$\hat{I}_{PM} = \sum_{i=1}^N f(m_i)(\theta_i - \theta_{i-1})$$

de manera que $\eta_i = m_i$ y $u_i = (\theta_i - \theta_{i-1})$.

En particular, si los puntos son equidistantes,

$$\hat{I}_{PM} = \frac{(b-a)}{N}\sum_{i=1}^{N} f\left(a + \frac{(2i-1)(b-a)}{2N}\right).$$

2) *Regla trapezoidal*. Está dada por

$$\hat{I}_T = \frac{1}{2}\sum_{i=1}^{N}\{f(\theta_i) + f(\theta_{i-1})\}(\theta_i - \theta_{i-1}),$$

por lo que en este caso $\eta_i = \theta_i$ y

$$u_i = \begin{cases} (\theta_1 - \theta_0)/2, & i = 0 \\ (\theta_{i+1} - \theta_{i-1})/2, & i = 1,\ldots N-1. \\ (\theta_N - \theta_{N-1})/2, & i = N \end{cases}$$

Si los nodos son equidistantes,

$$\hat{I}_T = \frac{(b-a)}{N}\left[\frac{1}{2}\{f(a) + f(b)\} + \sum_{i=1}^{N-1} f\left(a + \frac{i(b-a)}{N}\right)\right].$$

2.2 Métodos de Monte Carlo y simulación vía cadenas de Markov

- MÉTODOS DE MONTE CARLO: Los métodos de Monte Carlo permiten realizar aproximaciones de integrales mediante simulación. La idea básica consiste en expresar la integral requerida como el valor esperado de una función con respecto a alguna distribución de probabilidad. Existen varias formas de hacerlo, una de ellas es la siguiente.

- *Muestreo por importancia*. Supongamos que $f(\cdot)$ es una función real y se requiere evaluar la integral

$$I = \int_{\Theta} f(\theta) d\theta .$$

Claramente I también puede escribirse como

$$I = \int_{\Theta} \left\{ \frac{f(\theta)}{s(\theta)} \right\} s(\theta) d\theta = E_s \left\{ \frac{f(\theta)}{s(\theta)} \right\},$$

donde $s(\theta)$ es una función de densidad de probabilidad sobre Θ.

La distribución $s(\theta)$ se conoce como la distribución de muestreo por importancia y generalmente se elige de manera que sea fácil de simular.

Si generamos una muestra $\theta_1, \theta_2, ..., \theta_N$ de $s(\theta)$ entonces podemos aproximar la integral I a través del correspondiente momento muestral (estimador insesgado)

$$\hat{I}_{MI} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\theta_i)}{s(\theta_i)}.$$

- La precisión de $\hat{I}_{MI}$ depende tanto del tamaño de muestra N, como de $s(\theta)$.

Para producir una estimación con varianza pequeña se sugiere que $s(\theta)$:

a) sea fácil de simular,
b) tenga una forma similar a la de f(θ),
c) tenga la colas más pesadas que las de f(θ).

- MÉTODOS DE MONTE CARLO VIA CADENAS DE MARKOV (MCMC): Las técnicas de Monte Carlo vía cadenas de Markov permiten generar, de manera iterativa, observaciones de distribuciones multivariadas que difícilmente podrían simularse usando métodos directos.

- *La idea básica* consiste en construir una cadena de Markov que sea fácil de simular y cuya distribución de equilibrio corresponda a la distribución final que nos interesa.

- Teorema (Ergódico). *Sea* $\theta^{(1)}, \theta^{(2)},...,$ *una cadena de Markov homogénea, irreducible y aperiódica, con espacio de estados* Θ *y distribución de equilibrio* $f(\theta|\underline{x})$. *Entonces, conforme* $t\to\infty$,

  a) $\theta^{(t)} \xrightarrow{d} \theta$, *donde* $\theta \sim f(\theta|\underline{x})$;

  b) $\frac{1}{t}\sum_{i=1}^{t} g(\theta^{(i)}) \to E\{g(\theta)|\underline{x}\}$ *(Convergencia de promedios ergódicos).*

- *Algoritmo de Metropolis-Hastings*. Este algoritmo construye una cadena de Markov apropiada definiendo las probabilidades de transición de la siguiente manera:

  Sea $Q(\theta^*|\theta)$ una distribución auxiliar (arbitraria) y definamos

$$\alpha\left(\theta^*,\theta\right)=\min\left\{\frac{f\left(\theta^*|\underline{x}\right)Q\left(\theta|\theta^*\right)}{f\left(\theta|\underline{x}\right)Q\left(\theta^*|\theta\right)},1\right\}.$$

- Algoritmo. Dado un valor inicial $\theta^{(0)}$, la (t+1)-ésima iteración consiste en:
  1) Generar una observación θ* de $Q\left(\theta^*|\theta^{(t)}\right)$;
  2) Generar una variable u~U(0,1);
  3) Tomar $\theta^{(t+1)}=\begin{cases}\theta^*, \text{ si } u\le\alpha\left(\theta^*,\theta^{(t)}\right)\\ \theta^{(t)}, \text{ si } u>\alpha\left(\theta^*,\theta^{(t)}\right)\end{cases}$.

- Comentarios:
  a) Este procedimiento genera una cadena de Markov con distribución de transición $P\left(\theta^{(t+1)}|\theta^{(t)}\right)=\alpha\left(\theta^{(t+1)}|\theta^{(t)}\right)Q\left(\theta^{(t+1)}|\theta^{(t)}\right)$.
  b) La probabilidad de aceptación $\alpha\left(\theta^*,\theta\right)$, sólo depende de $f\left(\theta|\underline{x}\right)$ a través de un cociente, por lo que $f\left(\theta|\underline{x}\right)$ puede ser reemplazada por $f_x\left(\theta\right)=f\left(\underline{x}|\theta\right)f\left(\theta\right)$.

- Casos particulares de la distribución auxiliar $Q\left(\theta^*|\theta\right)$:
  1) Caminata aleatoria: $Q\left(\theta^*|\theta\right)=Q_1\left(\theta^*-\theta\right)$, donde Q_1 es una densidad de probabilidad simétrica centrada en el origen. Sugerencia: $Q\left(\theta^*|\theta\right)=N\left(\theta^*|\theta,\kappa V\left(\hat{\theta}\right)\right)$.
  2) Independencia: $Q\left(\theta^*|\theta\right)=Q_0\left(\theta^*\right)$, donde Q_0 es una densidad de probabilidad sobre Θ. Sugerencia: $Q_0\left(\theta^*\right)=N\left(\theta^*|\hat{\theta},\kappa V\left(\hat{\theta}\right)\right)$.

donde $\hat{\theta}$ y $V(\hat{\theta})$ son la media y la varianza de la distribución normal asintótica para $f(\theta|\underline{x})$, y $\kappa \geq 1$ es un factor de sobredispersión.

- *Muestreo de Gibbs*. Este algoritmo construye una cadena de Markov apropiada definiendo las probabilidades de transición a través de un proceso iterativo.

  Sea θ un vector de dimensión d, y sea $(\theta_1,...,\theta_k)$ una partición del vector θ, donde $\theta_i \in \Re_i^d$ y $\sum_{i=1}^{k} d_i = d$. Las densidades

  $$f(\theta_1|\theta_2,\ldots,\theta_k,\underline{x})$$
  $$\vdots$$
  $$f(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_k,\underline{x}),\ \ i=2,...,k-1$$
  $$\vdots$$
  $$f(\theta_k|\theta_1,\ldots,\theta_{k-1},\underline{x})$$

  se conocen como densidades condicionales completas y por lo general son fácil de obtenerse porque $f(\theta_i|\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_k,\underline{x}) \propto f(\theta|\underline{x})$ vista como función de θ_i.

  - Algoritmo. Dado un valor inicial $\theta^{(0)} = (\theta_1^{(0)},\ldots,\theta_k^{(0)})$, $\theta^{(t+1)}$ se obtiene de $\theta^{(t)}$ de la siguiente manera:

    1) Generar una observación $\theta_1^{(t+1)}$ de $f(\theta_1|\theta_2^{(t)},\theta_3^{(t)}\ldots,\theta_k^{(t)},\underline{x})$;
    2) Generar una observación $\theta_2^{(t+1)}$ de $f(\theta_2|\theta_1^{(t+1)},\theta_3^{(t)}\ldots,\theta_k^{(t)},\underline{x})$;

       $\vdots$

    3) Generar una observación $\theta_k^{(t+1)}$ de $f(\theta_k|\theta_1^{(t+1)},\theta_2^{(t+1)}\ldots,\theta_{k-1}^{(t+1)},\underline{x})$.

- Comentarios:
  a) La sucesión $\theta^{(1)},\theta^{(2)},...$ así obtenida es una realización de la cadena de Markov cuya distribución de transición está dada por

$$P\left(\theta^{(t+1)}\middle|\theta^{(t)}\right)=\prod_{i=1}^{k} f\left(\theta_i^{(t+1)}\middle|\theta_1^{(t+1)},\ldots,\theta_{i-1}^{(t+1)},\theta_{i+1}^{(t)},\ldots,\theta_k^{(t)},\underline{x}\right).$$

  b) En ocasiones la distribución final implica cierta estructura de independencia condicional entre algunos elementos del vector θ, por lo que muchas veces las densidades condicionales se simplifican.

- *Muestras*. Supongamos que se desea generar una muestra de tamaño N de la distribución $f\left(\theta|\underline{x}\right)$. Se tienen dos opciones:
  1) Correr N cadenas independientes: generar N valores iniciales $\theta_1^{(0)},\ldots\theta_N^{(0)}$ y correr una cadena de Markov con alguno de los dos algoritmos (Metrópolis-Hastings o Gibbs) y después de un cierto número de iteraciones T suficientemente grande, los valores $\theta_1^{(T)},\ldots\theta_N^{(T)}$ pueden considerarse como una muestra de tamaño N de la distribución final de θ.
  2) Correr una sola cadena: generar una sola cadena y tomar los valores $\theta^{(T+K)},\theta^{(T+2K)},\ldots,\theta^{(T+NK)}$ como una "muestra" de $f\left(\theta|\underline{x}\right)$, donde K se elige de manera que la correlación entre las observaciones sea pequeña.

     NOTA: Si K es pequeño, o incluso 1, $\theta^{(T+K)},\theta^{(T+2K)},\ldots,\theta^{(T+NK)}$ formarían un conjunto de observaciones dependientes y de acuerdo con el Teorema ergódico, serían suficientes para estimar cualquier valor esperado.

- *Convergencia*. En general no es fácil determinar en qué momento las cadenas han convergido. Un método empírico comúnmente utilizado, basado en el Teorema ergódico, consiste en graficar los promedios ergódicos de algunas funciones de θ contra el número de iteraciones y elegir el valor T a partir del cual las gráficas se estabilizan. En este caso es conveniente omitir los primeros valores de las cadenas. La idea de este período de calentamiento es permitir que las cadenas salgan de una primera fase de inestabilidad.

- EJERCICIOS DE CLASE: Utiliza el paquete WinBUGS para resolver los siguientes ejercicios:

  1) Sea θ la tasa de créditos hipotecarios otorgados por un banco. Durante el 2004 la tasa promedio fue de 60% y la desviación estándar de la tasa fue de 0.04. En lo que va del año 2005 se han solicitado 100 créditos, de los cuales se han otorgado únicamente 50.
     a) Usando la información del año pasado, encuentra la distribución beta que mejor describe el conocimiento inicial.
     b) Usando la información del año pasado, encuentra la distribución normal transformada que mejor describa el conocimiento inicial.
     c) Determina la distribución inicial de referencia.
     d) Usando los datos del año 2005 encuentra la distribución final para cada una de las distribuciones iniciales de los incisos (a) – (c).
     e) Estima la tasa de créditos otorgados, usando las 3 distribuciones finales del inciso (d).
     f) Estima el momio de otorgar un crédito, i.e., $\phi = \theta/(1-\theta)$, usando las 3 distribuciones finales del inciso (d).

2) Las utilidades mensuales de una compañía tienen una distribución $N(\mu,\sigma^2)$. Suponga que una muestra de 10 meses de esta compañía dio como resultado las siguientes utilidades: (212, 207, 210, 196, 223, 193, 196, 210, 202, 221).
   a) La incertidumbre sobre la utilidad promedio anual μ se puede representar por una distribución N(200,40), y la incertidumbre de la desviación estándar de las utilidades mensuales se puede representar mediante una distribución Ga(10,1). Mediante la distribución posterior estima μ y σ^2.
   b) Utilizando una distribución inicial no informativa, estima mediante la correspondiente distribución inicial μ y σ^2.

2.3 Medidas de comparación y ajuste de modelos

- Existen diversas medidas de bondad de ajuste y comparación de modelos. Alguna de ellas se basan en criterios predictivos, como las ordenadas predictivas condicionales (CPO), el logaritmo de la pseudo-verosimilitud marginal (LPML) y la medida-L. Otros indicadores se basan en comportamiento posterior de medidas de divergencia, como el criterio de información devianza (DIC).

- ORDENADAS PREDICTIVAS CONDICIONALES (CPO): La ordenada predictiva condicional es una estadística muy útil en la selección de modelos. Esta estadística fue propuesta originalmente por Geisser (1993) y Gelfand, Dey and Chang (1992).

  Para la i-ésima observación, la estadística CPO_i se define como

$$CPO_i = f\left(y_i \middle| D^{(-i)}\right) = \int f\left(y_i \middle| \theta, x_i\right) f\left(\theta \middle| D^{(-i)}\right) d\theta,$$

  donde y_i es la variable respuesta, x_i es el vector de covariables del individuo i y $D^{(-i)}$ denota los datos exceptuando el i-ésimo caso.

  CPO_i es la densidad predictiva posterior marginal de y_i dado $D^{(-i)}$ y se puede interpretar como la altura de esta densidad en la observación y_i. Por lo tanto valores grandes de CPO_i implican mejor ajuste del modelo.

- *Estimación Monte Carlo*: La estadística CPO puede ser aproximada de manera simple mediante técnicas Monte Carlo. Notemos primero que

$$CPO_i = \left(\int \frac{1}{f\left(y_i \middle| \theta, x_i\right)} f\left(\theta \middle| D\right) d\theta \right)^{-1}.$$

  Por lo tanto una aproximación Monte Carlo es de la forma

$$\hat{CPO}_i = \left(\frac{1}{R} \sum_{r=1}^{R} \frac{1}{f\left(y_i | \theta_r, x_i\right)} \right)^{-1},$$

donde $\theta_1, \ldots \theta_R$ es una muestra (MCMC) de la distribución posterior $f\left(\theta | D\right)$. Gráficas de CPO_i vs. i, o diagramas de caja son adecuados para comparar modelos.

- LOGARITMO DE LA PSEUDO-VEROSIMILITUD MARGINAL (LPML). Una forma de resumir las ordenadas predictivas condicionales es mediante una estadística resumen llamada logaritmo de la pseudo-verosimilitud marginal definida como:

$$LPML = \sum_{i=1}^{n} \log\left(CPO_i\right).$$

Valores *grandes* de LPML indican mejor ajuste.

- MEDIDA "L": La medida "L" es un método basado en un criterio de selección de modelos. Gelfand y Ghosh (1998) propusieron la siguiente medida

$$L(y) = \sum_{i=1}^{n} Var\left(Y_i^F | y\right) + \nu \sum_{i=1}^{n} \left\{ E\left(Y_i^F | y\right) - y_i \right\}^2,$$

donde Y_i^F es el valor predicho de y_i y $\nu \in (0,1)$ es un ponderador que determina un compromiso entre varianza y sesgo. Ibrahim, Chen y Sinha (1998) sugieren $\nu = 1/2$ para seleccionar el mejor modelo.

Valores *pequeños* de la medida "L" indican mejor ajuste.

- CRITERIO DE INFORMACIÓN DEVIANZA (DIC). Spiegelhalter et al. (2002) propuso una generalización del criterio de Akaike de comparación de modelos AIC. La generalización está basada en la distribución posterior de la devianza,

$$D(\theta) = -2\log f(y|\theta) + 2\log h(y),$$

donde $f(y|\theta)$ es la función de verosimiltud y $h(y)$ es una función de estandarización de los datos. Los autores sugieren resumir el ajuste del modelo por el valor esperado posterior de la devianza, $\overline{D} = E_{\theta|y}(D)$ y la complejidad del modelo por el número de efectivo de parámetros p_D. En al caso de modelos Gaussianos, se puede demostrar que una definición razonable para el número efectivo de parámetros es

$$p_D = E_{\theta|y}(D) - D(E_{\theta|y}(\theta)) = \overline{D} - D(\overline{\theta}).$$

Finalmente, el criterio de información devianza se define como

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\overline{\theta}).$$

- El criterio DIC puede ser estimado de manera simple mediante una muestra MCMC. Valores *pequeños* de DIC indican mejor ajuste.