CIMAT

Tarea 3

Inferencia Bayesiana

Javier Enrique Aguilar Romero

Abril 29, 2020

Sea el modelo de crecimiento logístico $\frac{dI}{dt} = \beta I(N - I)$ con $I(0) = I_0$. Este es el modelo de 0 compartimientos donde la velocidad de cambio en el numero de infecciosos I es proporcional al producto de I con el numero de susceptibles $S = N - I$; N es la población total susceptible. Este producto se puede ver como una aproximación de primer orden a la probabilidad de que un infeccioso se encuentre con un infectado. β es la intensidad de infección.

Suponga que tenemos conteos acumulados de casos confirmados de COVID19 y_i diarios que representan observaciones condicionalmente independientes para $I(t_i), t_1 < t_2 < ... < t_n$, esto es

$$y_i \sim Po\left(I(t_i) - I(t_{i-1})\right), i = 1, ..., n$$

La ODE de arriba tiene la siguiente solución analitica:

$$I(t) = \frac{N I_0 e^{\beta N t}}{N + I_0\left(e^{\beta N t} - 1\right)}.$$

El trabajo es hacer inferencia bayesiana para el parámetro β y la condición inicial I_0. Cual es la verosimilitud? Que a prioris pondría? Cual es la posterior?

En el ejercicio analizaremos los registros diarios acumulados para el Valle de México desde el 2 de Marzo hasta el 13 de abril. Tomaremos las siguientes convenciones y hechos

1. Los tiempos (dias) t comienzan en $t = 0$.

2. Para poder estudiar los casos diarios les denotamos por z_i en donde

$$\begin{aligned} z_0 &= y_0 \\ z_i &= y_i - y_{i-1}, \quad i = 1, .., n. \end{aligned}$$

   Por lo que los casos diarios son tales que

$$\begin{aligned} z_0 &\sim Po\left(I(t_0)\right) \\ z_i &\sim Po\left(I(t_i) - I(t_{i-1})\right), i = 1, ..., n \end{aligned}$$

   pues el texto menciona que y_i son casos acumulados; al hacer la diferencia de ellos tenemos los casos diarios.

3. Denotamos por J_i lo siguiente

$$J_0 = I(t_0)$$
$$J_i = I(t_i) - I(t_{i-1}), i = 1, ..., n$$

Por lo que $z_0 \sim Po(J_0)$ y $z_i \sim Po(J_i)$. La dependencia en t_i se sobreentiende.

4. $z_0, z_i, i = 1, .., n$ son tales que son condicionalmente independientes a I_0, β pues son función continua de funciones condicionalmente independientes a I_0, β.

El valor de N es la población del Valle de México que son 20 millones según [1]. Por lo que a priori esperamos que el valor de β sea sumamente pequeño. Dado que no tenemos mas información suponemos que $\beta \sim U(0, 1e-8)$ y para I_0 suponemos que los primeros infectados fueron entre 1 y 20 personas, es decir una uniforme discreta entre 1 y 20. Ademas suponemos independencia entre β e I_0. De esto se tiene que la priori $f(\beta, I_0)$ viene dada por

$$f(\beta, I_0) = \mathbb{1}_{(0,1e-8)}(\beta)\mathbb{1}_{\{1,...,20\}(I_0)}.$$

La verosimilitud $f(z_0, ..., z_n|\beta, I_0)$ tiene la forma

$$\begin{aligned} f(z_0, ..., z_n|\beta, I_0) &= \prod_{i=0}^{n} f(z_i|\beta, I_0) \\ &= \prod_{i=0}^{n} \frac{e^{-J_i} J_i^{z_i}}{z_i!} \\ &\propto e^{-\sum_{i=0}^{n} J_i} \prod_{i=0}^{n} J_i^{z_i}. \end{aligned}$$

Por lo que la posterior $f(\beta, I_0|z_0, ..., z_n)$ es proporcional a

$$f(\beta, I_0|z_0, ..., z_n) \propto e^{-\sum_{i=0}^{n} J_i} \prod_{i=0}^{n} J_i^{z_i} \mathbb{1}_{(0,1e-8)}(\beta)\mathbb{1}_{\{1,...,20\}(I_0)}$$

Para poder realizar la simulación se hizo uso de la implementación del t walk [2]. Se tienen 43 observaciones por lo que $n = 42$. Se realizaron 1 millón de iteraciones y se tomó un periodo de burn in de 50,000 iteraciones. Tomando un thinning de 150 se tenia una autocorrelación para de 0.048 para ambos parámetros. La cantidad de observaciones despues de esto es de 6334.

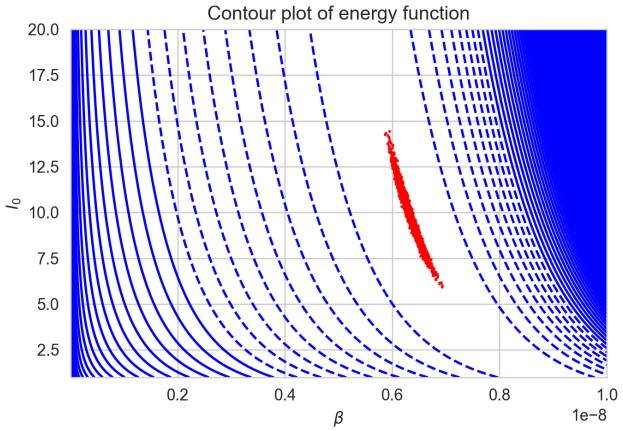
En la siguiente tabla θ representa el parámetro de interés, n_{eff} representa el effective sample size, también mostramos las varianzas de las observaciones de las cadenas y el estimador de la varianza de $\bar{\theta}$, dada por $\hat{V}(\theta) = \frac{\hat{V}(\bar{\theta})}{n_{eff}}$.

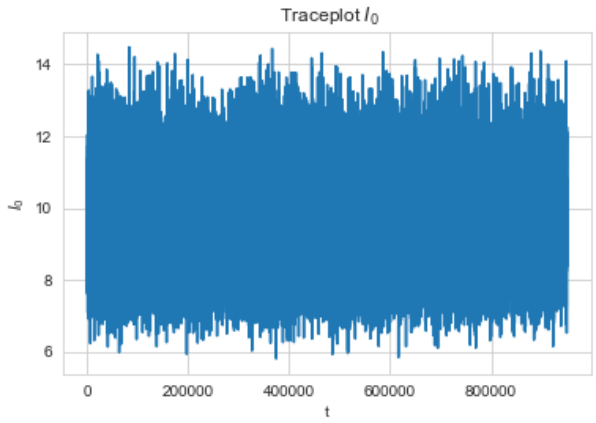
1. https://es.wikipedia.org/wiki/Zona_metropolitana_del_valle_de_M\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{e\global\mathchardef\accent@spacefactor\spacefactor}\accent19e\egroup\spacefactor\accent@spacefactorxico

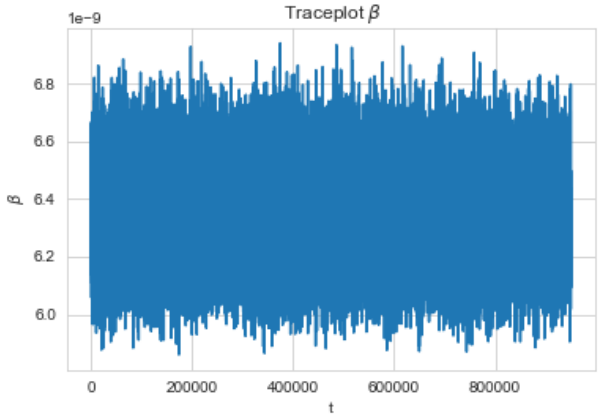
2. https://www.cimat.mx/~jac/twalk/

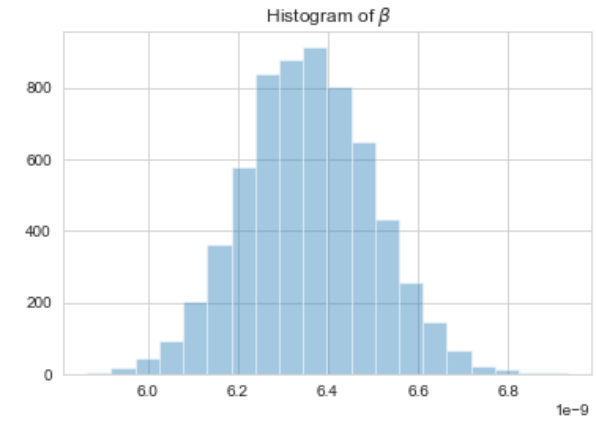
θ	n	n_{eff}	n_{eff}/n	$\bar{\theta}$	$\hat{V}(\theta)$	$\hat{V}(\bar{\theta})$
β	6334	5704	0.90	6.35e-09	2.08e-20	3.65e-24
I_0	6443	5666	0.89	9.675	1.425	0.00025

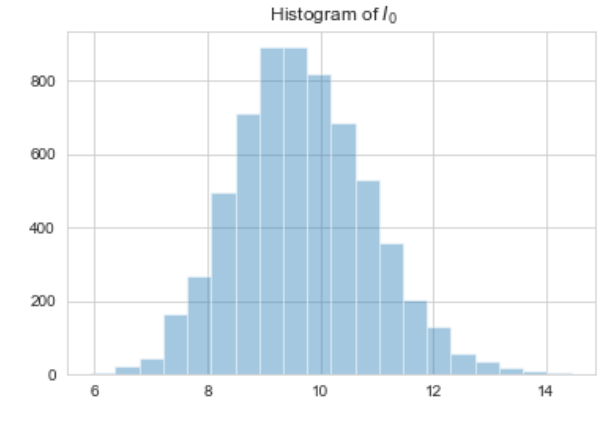
En el siguiente gráfico se presenta la traza tras el burn in y sin el thinning, así como el histograma de las 6334 observaciones finales. Tambien se presentan los contornos de la función de energía y las muestras finales tomadas.











La siguiente tabla muestra un intervalo de credibilidad del 95% para θ así como la media posterior.

θ	$q_{0.025}$	$\bar{\theta}$	$q_{0.975}$
β	6.077e-09	6.35e-09	6.644e-09
I_0	7.493	9.675	12.136

Tomando como estimador puntual la media posterior podemos generar valores ajustados de la siguiente forma. Se considera la curva $\hat{I}(t)$, que es tomar $I(t)$ con los valores $\bar{\beta}, \bar{I}_0$ como parámetros. Luego se generan valores ajustados $\hat{z}_i \sim Po\left(\hat{I}(t_i) - \hat{I}(t_{i-1})\right)$. Una medida de evaluación puede ser la suma de cuadrados del error, es decir SSE $= \sum_{i=0}^{N}(z_i - \hat{z}_i)^2 = 28864$.

En los siguientes gráficos se presentan: el ajuste a los casos diarios y el ajuste a casos acumulados. En ambos casos se presentan bandas de credibilidad del 95%. Dichas bandas fueron formadas tomando valores ajustados con $q_{0.025}, q_{0.0975}$.