Universidad Nacional de Rosario

LICENCIATURA EN ESTADÍSTICA

UNA MASACRE ESTADÍSTICA

Regresión Lineal Bayesiana

Autores: Franco Santini - Nicolas Gamboa - Andrés Roncaglia

Docentes: Ignacio Evangelista - Tomás Capretto

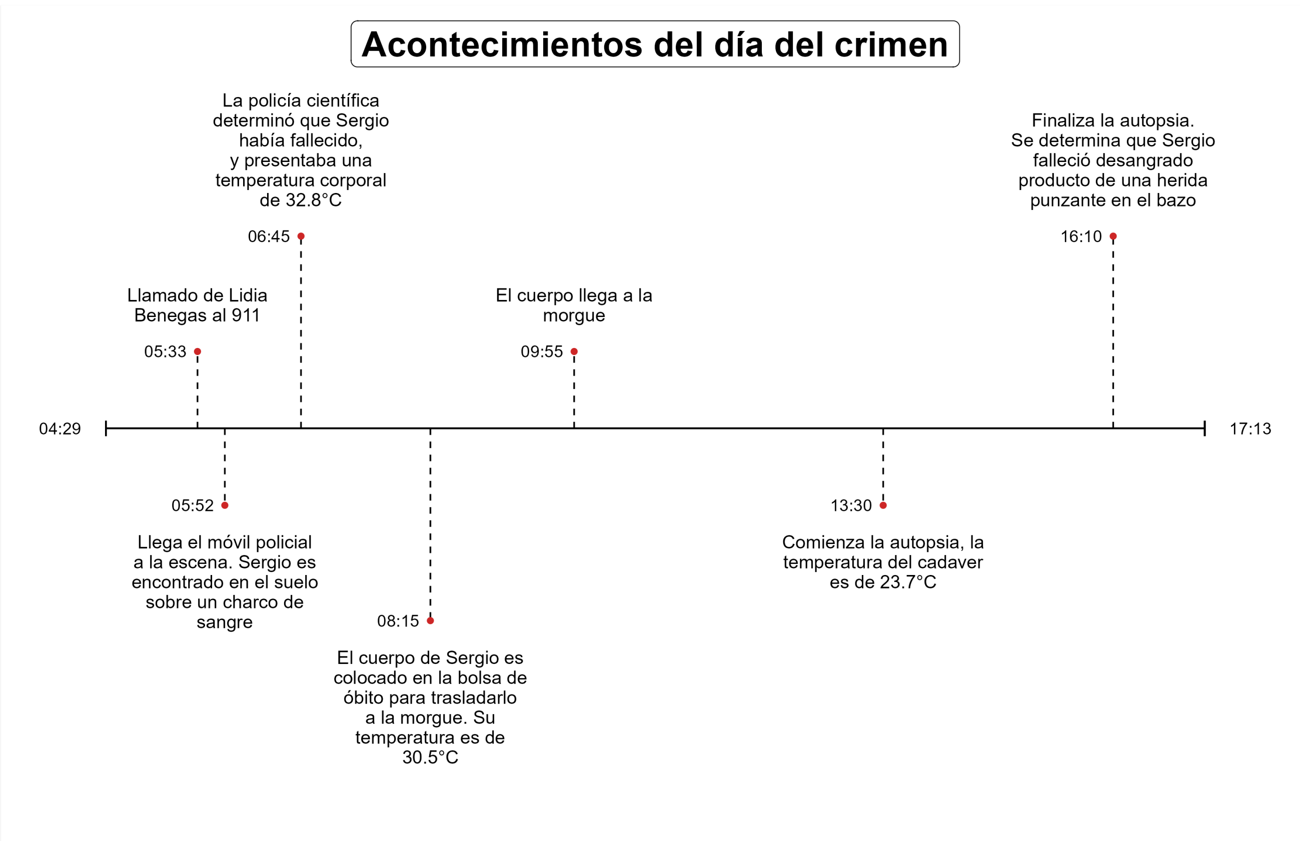
2024

Tabla de contenidos

Introducción

Siguiendo los acontecimientos de la muerte de Sergio, una templada madrugada de miércoles en Salsipuedes - Córdoba, la policía forense busca determinar la hora de su muerte. El encargado de esto es Guido, un estudiante de estadística fascinado por poder al fin aplicar sus conocimientos estadísticos en un caso de criminalística.

Los datos que la policía foresense propició a Guido son los siguientes:



Metodología

En la medicina forense existen muchas maneras de estimar la hora de la muerte. Entre ellas, una de las más utilizadas es el estudio de la temperatura corporal.

Se sabe que la temperatura de un cuerpo satisface la siguiente ley:

$$\frac{dT(t)}{dt} = r[T_{amb} - T(t)]$$

Donde T_{amb} es la temperatura ambiente en donde se encuentra el cuerpo, en este caso 22 °C, r es una constante que determina el ritmo de enfriamiento y $T(t)$ es la función que describe la temperatura del cadaver en función del tiempo después de la muerte.

Una posible función $T(t)$ que satisface esta ecuación diferencial es la siguiente:

$$
\begin{aligned}
T(t) = T_{\text{amb}} + (T_i - T_{\text{amb}})e^{-rt} \Rightarrow \frac{dT(t)}{dt} = -r(T_i - T_{\text{amb}}) \cdot e^{-rt} \Rightarrow \\
\Rightarrow r(T_{\text{amb}} - T_{\text{amb}}) \cdot e^{-rt} - r(T_i - T_{\text{amb}}) \cdot e^{-rt} = r\{T_{\text{amb}} - T_{\text{amb}} - (T_i - T_{\text{amb}})e^{-rt}\} = \\
= r\{T_{\text{amb}} - \underbrace{[T_{\text{amb}} + (T_i - T_{\text{amb}})e^{-rt}]}_{T(t)}\} = r[T_{\text{amb}} - T(t)]
\end{aligned}
$$

Algunos valores posibles para la constante de enfriamiento r son:

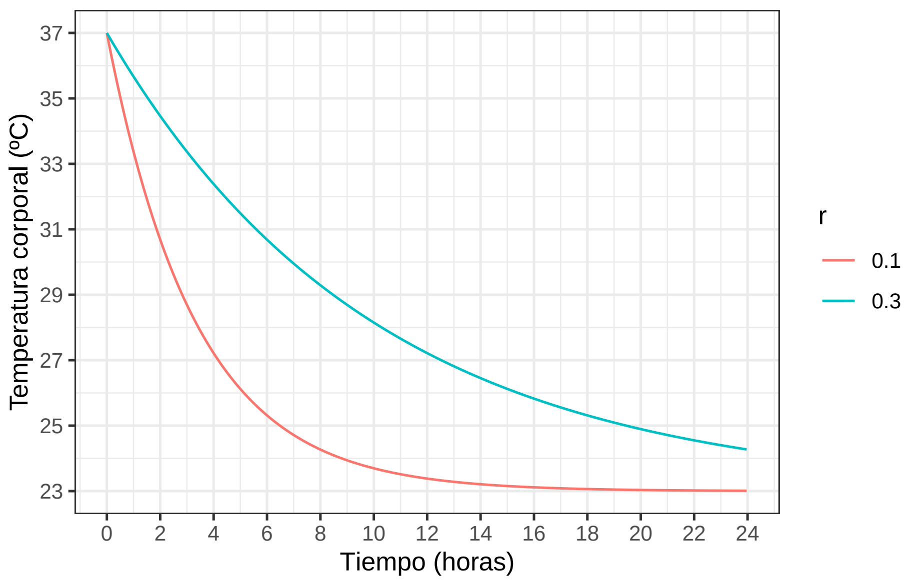


Figura 1: Temperatura del cuerpo según tiempo post-mortem. Temperatura ambiente: 23 °C

Al ser su primer caso oficial, Guido se sintió un poco desconcertado por no saber a qué hora podría haber ocurrido la tragedia. Es por esto que pasó días y días investigando (sin dormir) cuanto tiempo tarda en enfriarse un cuerpo luego de que la persona muera. Finalmente, encontro lo que buscaba, una publicación de La Gaceta Internacional de Ciencias Forenses, en la cuál se detalla que, la tasa de enfriamiento del cuerpo varía en tres estapas, las primeras 3 o 4 horas donde la temperatura baja menos de medio grado por hora, las siguientes 6 a 10 horas donde la temperatura baja alrededor de 1 grado por hora, y a partir de las 10 horas el cuerpo desciende su temperatura entre 3/4 a 1/4 grados por hora, hasta alcanzar la temperatura ambiente. Teniendo en cuenta que r se supone constante, se puede asumir que un buen valor de r es mayor que 0,1 y menor que 0,45.

Con el objetivo de finalmente, después de tanto esfuerzo, estimar la hora de la muerte de Sergio, Guido decide utilizar un enfoque Bayesiano, planteando el siguiente modelo lineal.

$$
T(t) - T_{\text{amb}} = T_{\text{diff}}e^{-rt} \Rightarrow ln(T(t) - T_{\text{amb}}) = ln(T_{\text{diff}}e^{-rt}) \Rightarrow \underbrace{ln(T_{\text{diff}})}_{\beta_0} - \underbrace{r}_{\beta_1} t = \beta_0 - \beta_1 \cdot t
$$

Donde β_0 representa el logaritmo natural de la diferencia entre la temperatura inicial del cuerpo y la temperatura ambiente y β_1 es el cambio esperado en el logaritmo de la diferencia de temperaturas por cada hora.

Una transformación posible para este modelo, la cuál permite interpretar a t como el tiempo después de haber encontrado el cuerpo, se obtiene mediante una traslación utilizando un nuevo parámetro δ. Resultando:

$$T(t+\delta)-T_{\text{amb}} = T_{\text{diff}}e^{-r(t+\delta)} \Rightarrow \Rightarrow ln(T(t+\delta)-T_{\text{amb}}) = ln(T_{\text{diff}}e^{-r(t+\delta)}) \Rightarrow \Rightarrow \underbrace{ln(T_{\text{diff}})}_{\beta_0} - \underbrace{r}_{\beta_1}(t+\delta) = \beta_0-\beta_1\cdot(t+\delta)$$

El modelo que se usa para estimar la hora de muerte de Sergio resulta:

$$ln(T(t+\delta)-T_{\text{amb}}) = \beta_0 - \beta_1 \cdot (t+\delta) \qquad (1)$$

Donde:

- δ es el tiempo (en horas) que transcurrió entre la que persona murió y se encontró el cuerpo.
- Los demás parámetros conservan su significado.

Guido recuerda haber mirado el pronóstico del tiempo el día de la muerte de Sergio y recuerda que anunciaban una temperatura ambiente de 22 ºC. Por otro lado, dado que la temperatura corporal de una persona sana ronda los 37 °C, consideró esta la temperatura inicial del cuerpo de Sergio.

$$ln(T(0)-T_{amb}) = \beta_0 - \beta_1 \cdot (0) \Rightarrow \beta_0 = ln(37-22) = ln(15) = 2.7081$$

Al ser un enfoque bayesiano, Guido debe pensar una distribución a priori para los parámetros δ y β_1.

En principio, dado que el cuerpo fue encontrado con una temperatura corporal de 32.8 °C, y sabiendo que Sergio murió desangrado, lo cual acelera considerablemente el descenso de la temperatura, cree que debe haber muerto hace aproximadamente 2 horas, aunque no está para nada seguro, por lo que utilizará una distribución uniforme, que irá entre 1 hora y media y 2 horas y media.

Con respecto al parámetro β_1, considerando la hora a la que estima que murió Sergio y la temperatura a la que fue encontrado, observa lo siguiente.

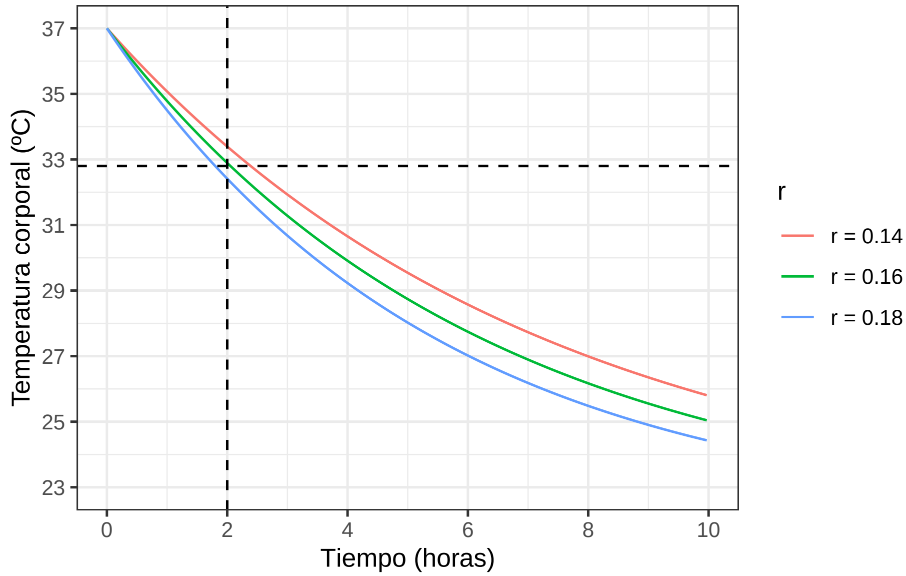


Figura 2: Temperatura del cuerpo según tiempo post-mortem. Temperatura ambiente: 22 °C

En base a esto, tomará valores similares a los graficados para la distribución de β_1, considerando una distribución normal.

Con respecto al parámetro σ, cree que un termómetro puede fallar a lo sumo en 2 °C, por lo que postula una distribución media normal, dado que σ no puede tomar valores negativos, con un desvío de 0.25.

De manera un poco más formal, el modelo planteado por Guido es el siguiente, suponiendo que los parámetros, a priori, son independientes.

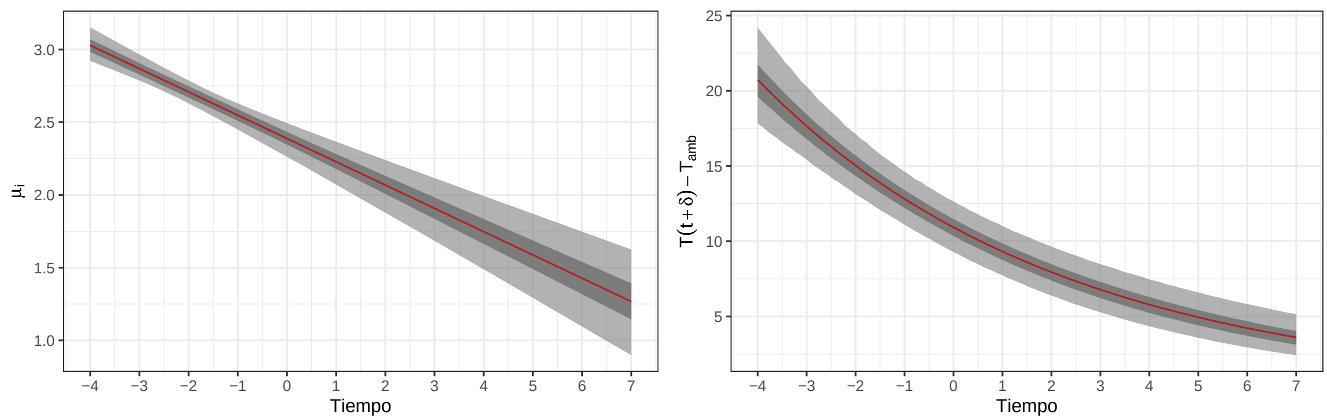
$$
\begin{aligned}
ln(Y)/\mu_i,\ \sigma &\sim N(\mu_i,\ \sigma) \\
\mu_i &= \beta_0 - \beta_1 \cdot (t + \delta) \\
\beta_0 &= 2.7081 \\
\beta_1 &\sim N^+(0.16,\ 0.02) \\
\delta &\sim U(1.5,\ 2.5) \\
\sigma &\sim N^+(0,\ 0.25)
\end{aligned}
$$

El interés se centra principalmente en el parámetro δ

Resultados

El objetivo de Guido es estimar la hora de la muerte y para hacerlo usará toda la información que tenga a su disposición. Pero antes de esto, quiere asegurarse de que sus creencias iniciales tengan sentido, por lo que decide realizar pruebas predictivas a priori.

Pruebas predictivas a Priori



(a) Del predictor lineal

(b) De $T(t) - T_{amb}$

Figura 3: Distribución a Priori

Sumado a esto decide buscar un intervalo de tiempo en el que pueda decir con un 90% de certeza que Sergio murió dentro de ese rango horario.

A priori, Guido concluye con una credibilidad del 90% que Sergio murió entre las 04:18hs y las 05:12hs, por lo que las distribuciones a priori que planteó parecen tener sentido.

Pruebas predictivas a Posteriori

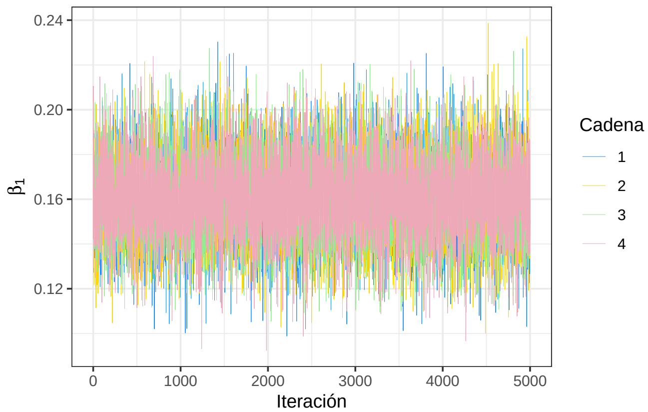
Ahora sí, finalmente Guido utilizará los datos provistos para estimar la hora de la muerte, pero con el objetivo de desafiarse a sí mismo decide hacerlo de manera secuencial, para ver como los datos van afectando sus estimaciones.

Los datos disponibles son los siguientes:

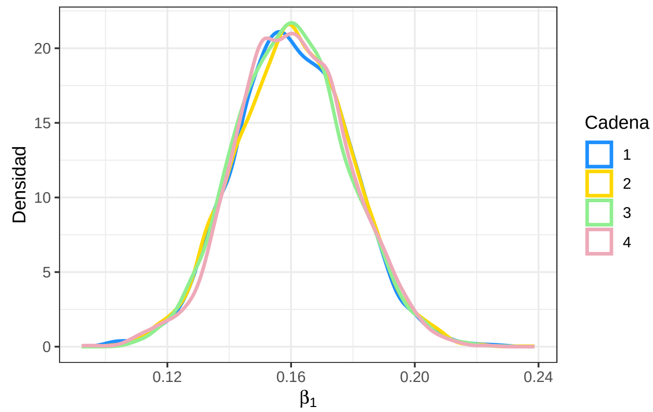
$Tiempo$	$Temp_{diff}$	$ln(Temp_{diff})$
0.00	10.8	2.380
1.50	8.5	2.140
6.75	1.7	0.531

Después de que la computadora de bajos recursos del departamento de la Policía Científica estuviera encendida durante días, Guido consiguió correr los modelos en Stan.

A Guido le remarcaban mucho en clases de Estadística Bayesiana la importancia de verificar la convergencia de las cadenas de Markov, en este caso obtenidas por el método NUTS (*No-U-Turn Sampler*), para los distintos párametros del modelo. Es por ello que para cada una de las estimaciones de los parámetros, presenta gráficos de densidad y de traza.

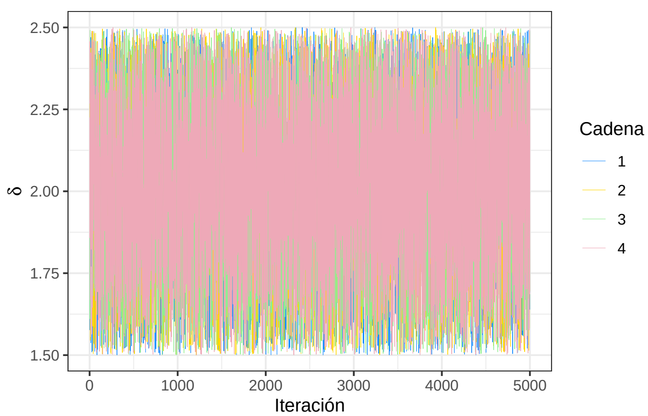


(a) Traza

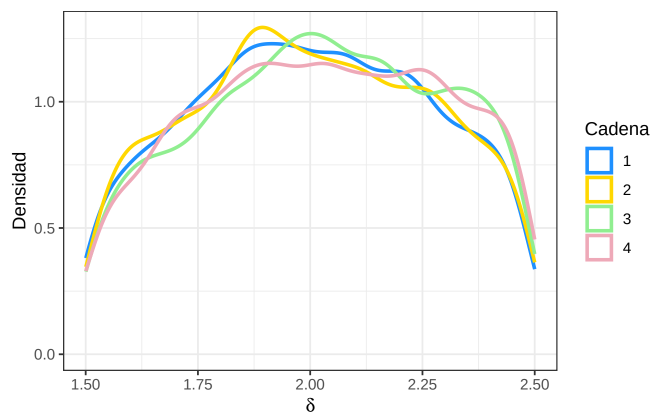


(b) Densidad

Figura 4: Cadenas de Markov para β_1

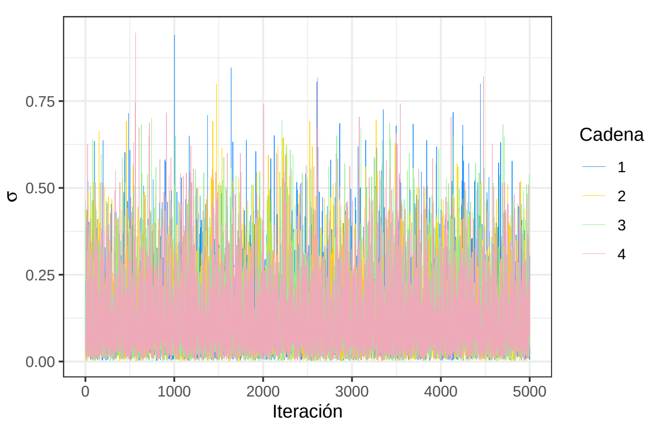


(a) Traza

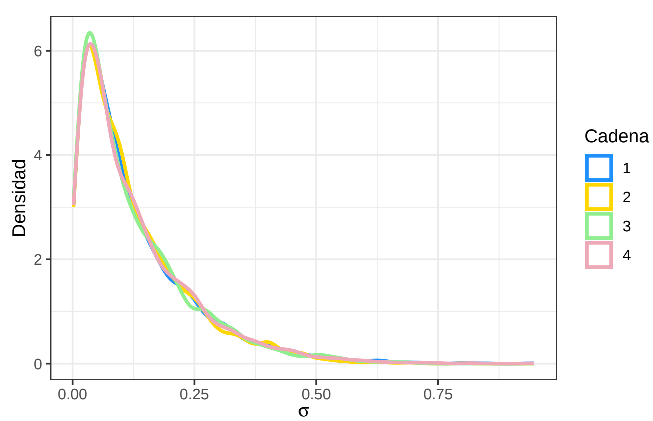


(b) Densidad

Figura 5: Cadenas de Markov para δ



(a) Traza

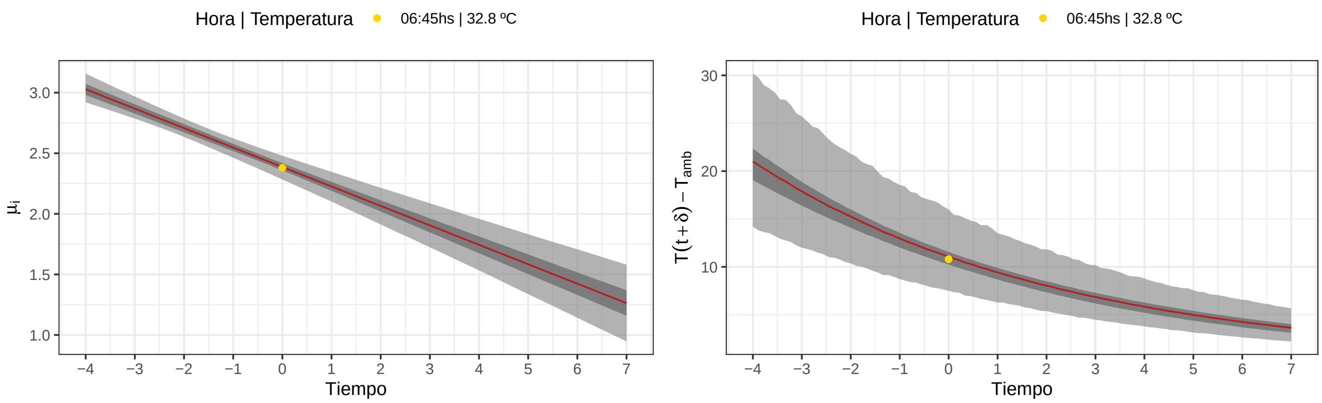


(b) Densidad

Figura 6: Cadenas de Markov para σ

Con la información provista hasta las 7 de la mañana, las cadenas de Markov para cada uno de los parametros parecen converger a la misma distribucion a posteriori. Esto, además de haber obtenido un $\hat{R} = 1$ para todos los párametros, le confirman a Guido que las muestras obtenidas son representativas.

La curiosidad de Guido lo llevaba a querer visualizar como los datos afectan al prior, por lo que graficó bandas de credibilidad para el predictor lineal y para la diferencias de temperaturas.



(a) Del predictor lineal

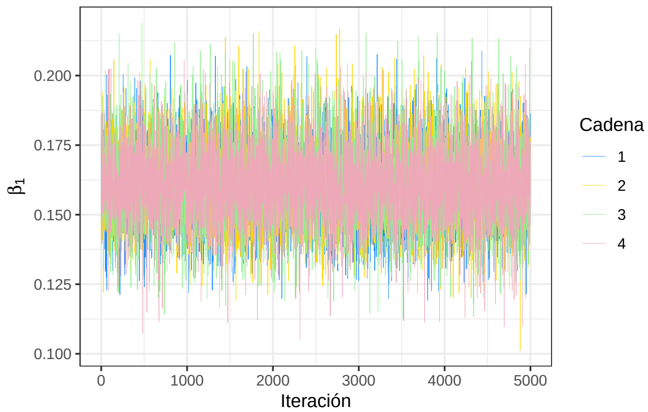
(b) De $T(t) - T_{amb}$

Figura 7: Distribuciónes a Priori con la información hasta las 7hs

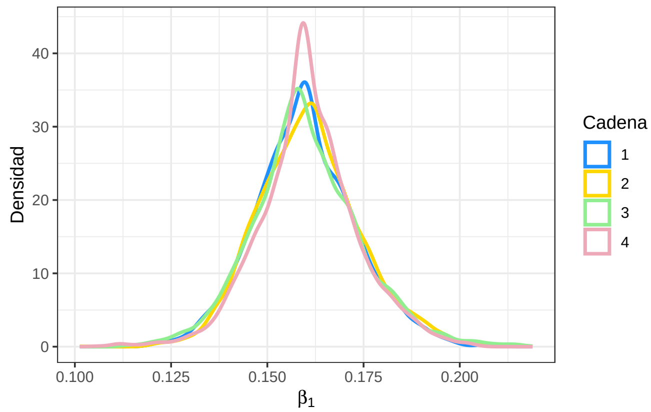
Guido observa que esta primera observación parece acotar levemente el intervalo de tiempo en el que puede decir con una buena certeza que Sergio murió. Ahora él, puede decir con la misma credibilidad de antes que Sergio falleció entre las 04:19hs y las 05:11hs, calculado a partir de la distribución a posteriori de δ, que se muestra en la Figura 5. Esto no parece mucho, pero le sirve a Guido para saber que va por buen camino, lo cuál lo motiva a seguir.

Guido procede a seguir añadiendo datos a su modelo, con lo que usa la información disponible hasta las 10hs.

De igual forma que con un solo dato, todos los $\hat{R}$ son iguales a 1, indicando que las cadenas convergen a la misma distribución para todos los parámetros. Esto se puede ver en los siguientes gráficos de traza y densidad para cada uno de los parámetros.

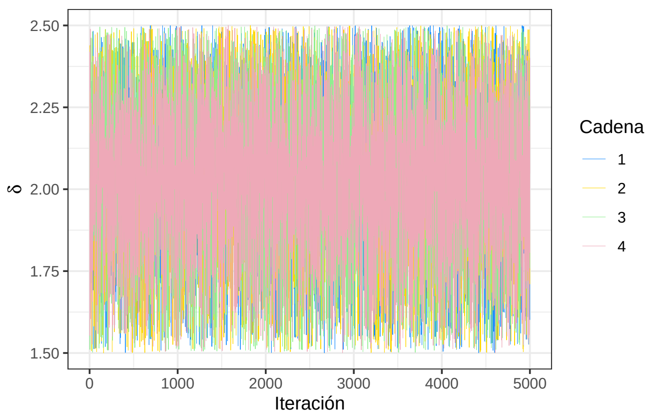


(a) Traza

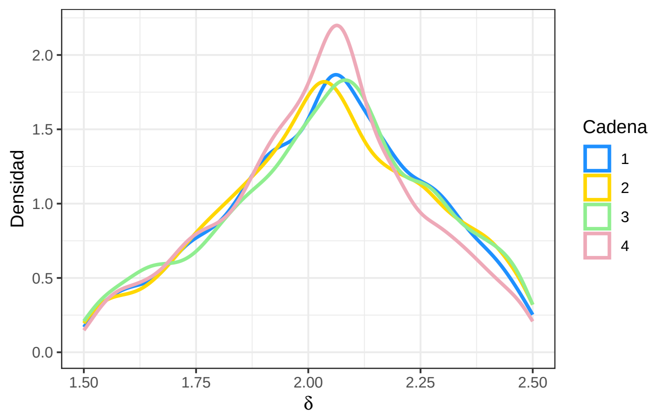


(b) Densidad

Figura 8: Cadenas de Markov para β_1

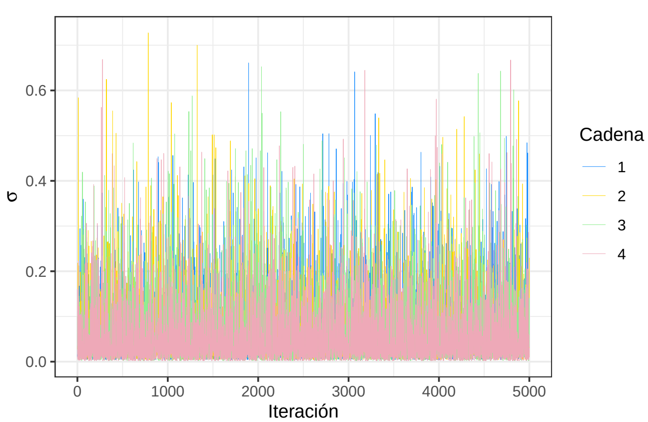


(a) Traza

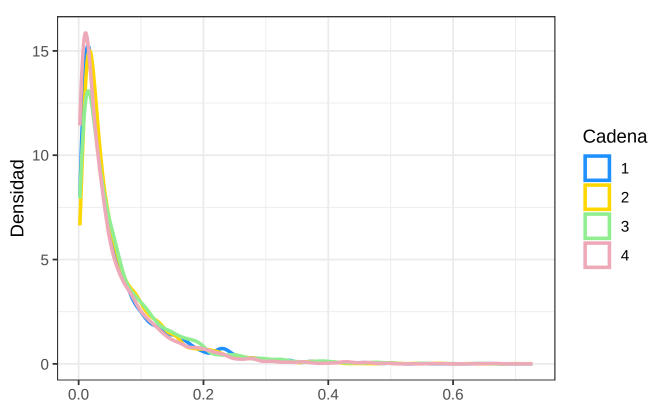


(b) Densidad

Figura 9: Cadenas de Markov para δ

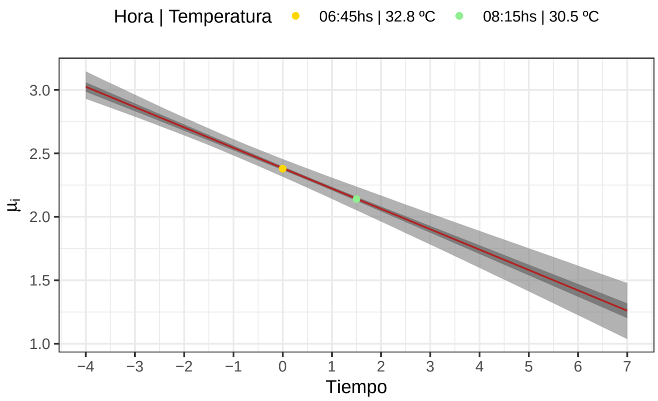


(a) Traza

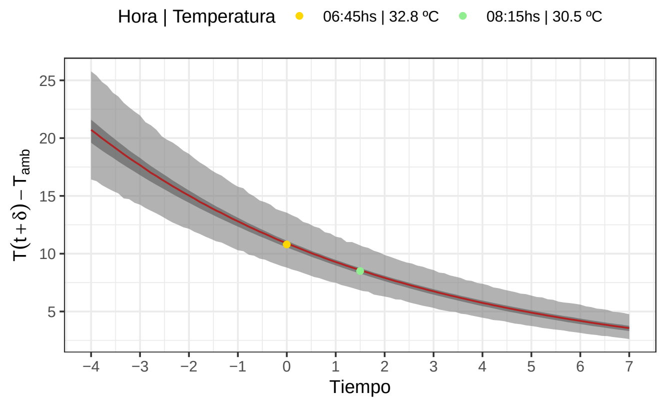


(b) Densidad

Figura 10: Cadenas de Markov para σ



(a) Del predictor lineal

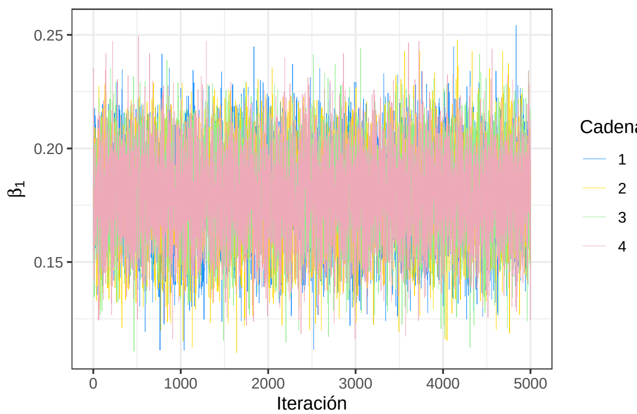


(b) De $T(t) - T_{amb}$

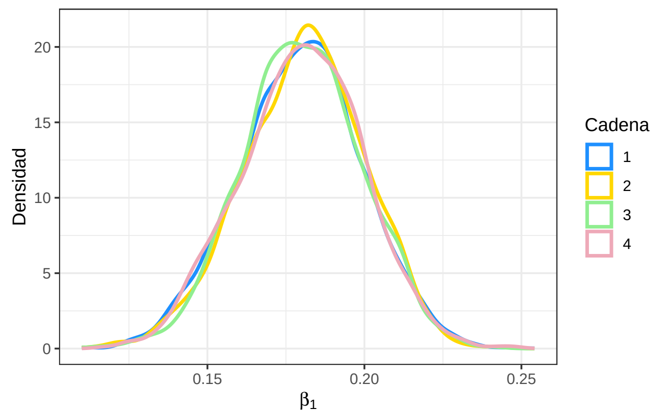
Figura 11: Distribuciónes a Priori con la información hasta las 10hs

Guido ahora concluye con una certeza del 90% que Sergio pereció entre las 04:20hs y las 05:07hs, calculado a partir de la distribución a posteriori de δ, que se muestra en la Figura 9.

Finalmente, decide hacer uso de toda la información disponible, aunque la última temperatura registrada del cuerpo le traía un mal presentimiento...

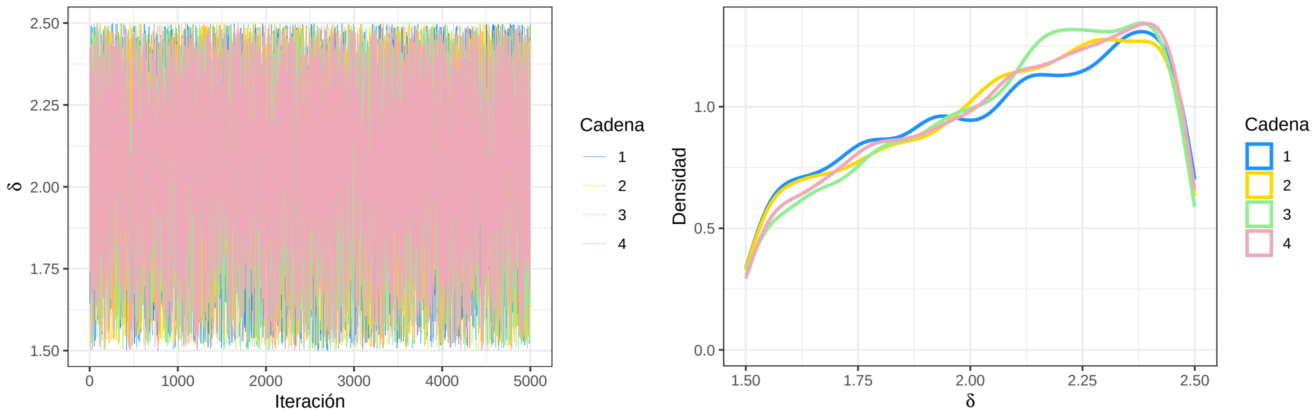


(a) Traza



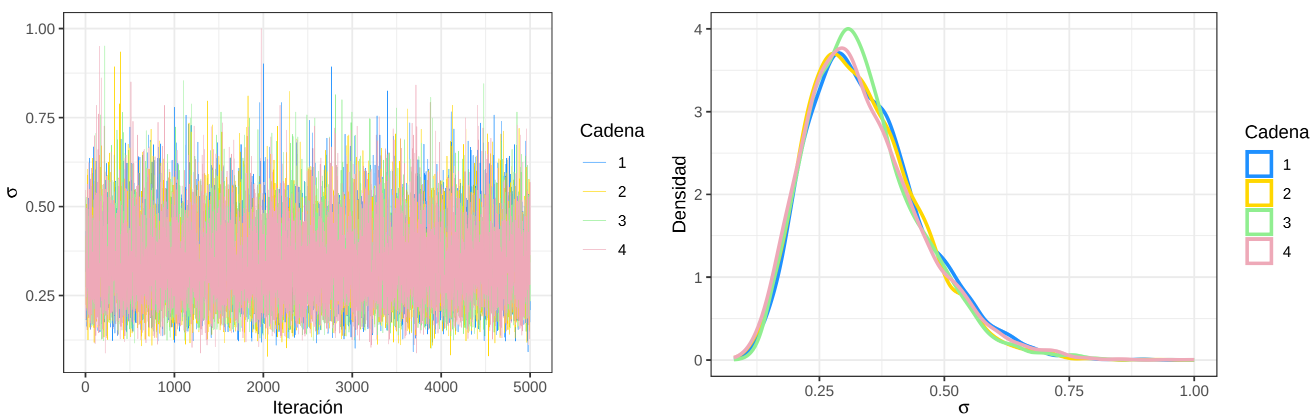
(b) Densidad

Figura 12: Cadenas de Markov para β_1



(a) Traza (b) Densidad

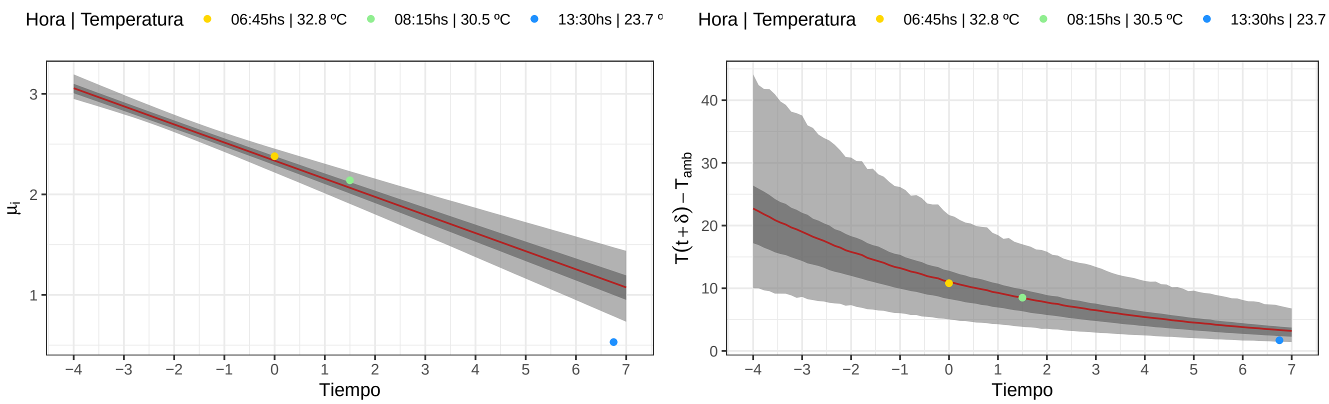
Figura 13: Cadenas de Markov para δ



(a) Traza (b) Densidad

Figura 14: Cadenas de Markov para σ

Hasta el momento, todo iba bien, las cadenas aún convergían con todos los $\hat{R}$ iguales a 1, pero al ver en la Figura 13 la densidad estimada del parámetro δ sospechaba que algo no encajaba. Sin embargo, decidió seguir adelante.



(a) Del predictor lineal

(b) De $T(t) - T_{amb}$

Figura 15: Distribuciónes a Priori con toda la información disponible

Con estos resultados, Sergio pudo haber fenecido entre las 04:17hs y las 05:10hs, un intervalo de tiempo mayor que el que se había obtenido anteriormente con la misma credibilidad, calculado a partir de la distribución a posteriori de δ, que se muestra en la Figura 13.

"¡Lo que sospechaba!" dijo Guido, al asimilar que la última temperatura registrada de Sergio fue en la morgue, un lugar que tiende a ser frío para la preservación de los cadáveres, con lo que la temperatura del cuerpo de Sergio pudo haberse enfriado más rapidamente que en el lugar del hecho, donde recuerda que había una temperatura ambiente de 22 ºC.

Pero esto no era todo, pensó Guido, ¿Qué hubiera pasado si Sergio no gozaba de un buen estado de salud?. Por lo que, si tuviese información que indique que Sergio no gozaba de buena salud al momento de fallecer, esto podría modificar su creencia a priori, ya que la velocidad de enfriamiento se retarda en casos de enfermedades febriles, personas con obesidad e intoxicaciones por nicotina, rodenticidas, entre otros. También la perdida de calor del cuerpo puede acelerarse en personas con bajo peso, enfermedades crónicas e intoxicadas por fósforo, arsénico y alcohol[1].

Una posible solución que pensó, podría ser agregar una variable que indique si la persona gozaba de un buen estado de salud, junto a un parámetro β_2 que modifique la pendiente original. Resultando el modelo:

$$ln(T(t) - T_{amb}) = \underbrace{ln(T_{\text{diff}})}_{\beta_0} - \underbrace{r}_{\beta_1} \cdot (t+\delta) + \beta_2 \cdot I \cdot (t+\delta) = \beta_0 - \beta_1 \cdot (t+\delta) + \beta_2 \cdot I \cdot (t+\delta) = \beta_0 + (\beta_2 \cdot I - \beta_1) \cdot (t+\delta)$$

Donde:

- Las interpretaciones de β_0, β_1 y δ son las mismas que el modelo (1).
- β_2: es la diferencia entre la tasa de enfriamiento de una persona que goza de un buen estado de salud de la que no lo goza.
- $I = \begin{cases} 0 & \text{si Sergio gozaba de un buen estado de salud} \\ 1 & \text{si Sergio no gozaba de un buen estado de salud} \end{cases}$

[1] La Gaceta Internacional de Ciencias Forenses

Conclusión

Guido decide informarle a los policías que Sergio murió entre las 04:20hs y las 05:07hs, considerando a las 04:43hs como la hora más probable de la muerte, ignorando el último dato, dado que consideró que este empeoraba las estimaciones por haberse obtenido en condiciones diferentes a las del lugar del crimen. La policía forense agradece a Guido por su esfuerzo y se pondrán a investigar más detenidamente el caso con la información brindada. Guido antes de despedirse le recomendó a los policias verificar el repositorio donde realizó toda su investigación.