Análisis Bayesiano del Modelo de Regresión Lineal

Manuel Mendoza R.

Departamento de Estadística
Instituto Tecnológico Autónomo de México

IV Escuela de Verano. Centro de Estadística Aplicada a Estudios Socioeconómicos.
Medellín, Colombia. Diciembre 5-7, 2011.

Contenido

- Formulación del modelo
- Análisis Bayesiano
- El modelo Normal Lineal
- Análisis Conjugado
- Iniciales Mínimo Informativas
- Ilustración
- Comentarios

Contenido

- Formulación del modelo
- Análisis Bayesiano
- El modelo Normal Lineal
- Análisis Conjugado
- Iniciales Mínimo Informativas
- Ilustración
- Comentarios

Formulación del modelo

- *Objetivo*: Describir la variable aleatoria Y, con soporte $\mathcal{Y}$, en las condiciones de observación $\underline{X}$ a través de la función de probabilidad *condicional* $P(Y \mid \underline{X}, \theta)$.

  - ✓ La distribución se considera totalmente conocida, excepto por valor del parámetro fijo, de dimensión finita, $\underline{\theta}$.

  - ✓ Se cuenta con una colección de observaciones

$$(Y_i, \underline{X}_i); \; i = 1, 2, \ldots, n$$

Formulación del modelo

- Las observaciones $(Y_i, \underline{X}_i)$; $i = 1, 2, \ldots, n$, son tales que:

  - ✓ Las variables Y_i ; $i = 1, 2, \ldots, n$, son condicionalmente independientes, dados $\underline{\theta}$ y $\underline{X}_i$; $i = 1, 2, \ldots, n$.

$$P(Y_{(n\times 1)} \mid X_{(n\times k)}, \underline{\theta}) = \prod P(Y_i, \mid \underline{X}_i, \underline{\theta})$$

  - ✓ Antes de los datos $(Y_{(n\times 1)}, X_{(n\times k)})$, la información sobre $\underline{\theta}$ se describe con la probabilidad inicial (*a priori*) $P(\underline{\theta})$.

Análisis Bayesiano

- ✓ Desde la perspectiva Bayesiana, sabemos que el fenómeno bajo estudio se aborda con el *modelo conjunto*

$$P(\underline{Y}_{(n\times 1)},\ X_{(n\times k)}, \underline{\theta})$$

que describe la información disponible, tanto sobre los datos $Y_{(n\times 1)}$, $X_{(n\times k)}$ como sobre el parámetro $\underline{\theta}$.

- ✓ Este modelo se puede representar como:

$$P(Y_{(n\times 1)},\ X_{(n\times k)}, \underline{\theta}) = P(Y_{(n\times 1)} \mid X_{(n\times k)}, \underline{\theta})\ P(X_{(n\times k)} , \underline{\theta})$$

$X_{(n\times k)}$ es como otro parámetro!

Análisis Bayesiano

- ✓ Así, permite describir a Y, a partir de covariables y parámetros.

$$P(\underline{Y}_{(n\times 1)},\ X_{(n\times k)},\ \underline{\theta}) = P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)},\ \underline{\theta})\ P(X_{(n\times k)},\ \underline{\theta})$$

- ✓ Otra factorización produce:

$$P(\underline{Y}_{(n\times 1)},\ X_{(n\times k)},\ \underline{\theta}) = P(\underline{Y}_{(n\times 1)},\ X_{(n\times k)} \mid \underline{\theta})\ P(\underline{\theta})$$

de donde es claro que $\underline{\theta}$ representa los parámetros de la distribución *conjunta* de $\underline{Y}_{(n\times 1)}$ y $X_{(n\times k)}$.

Análisis Bayesiano

- Entonces, si $\underline{\theta}^t = (\underline{\phi}, \underline{\gamma})$ donde $\underline{\phi}$ es el parámetro de la distribución condicional de Y, mientras $\underline{\gamma}$ es el parámetro de la distribución marginal de X, se tiene

$$P(\underline{Y}_{(n\times1)}, \ X_{(n\times k)}, \underline{\theta}) = P(\underline{Y}_{(n\times1)} \mid X_{(n\times k)}, \underline{\theta}) \, P(X_{(n\times k)} , \underline{\theta})$$

$$= P(\underline{Y}_{(n\times1)} \mid X_{(n\times k)}, \underline{\phi}) \, P(X_{(n\times k)} , \underline{\theta})$$

- Pero, además,

$$P(X_{(n\times k)}, \underline{\theta}) = P(X_{(n\times k)} \mid \underline{\theta}) \, P(\underline{\theta})$$

$$= P(X_{(n\times k)} \mid \underline{\gamma}) \, P(\underline{\phi}, \underline{\gamma})$$

$\underline{\theta}^t = (\underline{\phi} , \underline{\gamma})$

Análisis Bayesiano

- ✓ En resumen, se obtiene:

$$P(\underline{Y}_{(n\times 1)},\ X_{(n\times k)},\ \underline{\theta}) = P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)},\ \underline{\phi})\ P(X_{(n\times k)} \mid \underline{\gamma})\ P(\underline{\phi},\ \underline{\gamma})$$

- ✓ Bajo condiciones generales es razonable suponer que los parámetros $\underline{\phi}$ y $\underline{\gamma}$ son independientes y entonces,

$$P(\underline{Y}_{(n\times 1)},\ X_{(n\times k)},\ \underline{\theta}) = P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)},\ \underline{\phi})\ P(X_{(n\times k)} \mid \underline{\gamma})\ P(\underline{\gamma})\ P(\underline{\phi})$$

$$= \left(P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)},\ \underline{\phi})\ P(\underline{\phi}) \right) \times \left(P(X_{(n\times k)} \mid \underline{\gamma})\ P(\underline{\gamma}) \right)$$

Análisis Bayesiano

$$P(\underline{Y}_{(n\times 1)},\ X_{(n\times k)},\ \underline{\theta}) = \left(P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)},\ \underline{\phi})\, P(\underline{\phi}) \right) \times \left(P(X_{(n\times k)} \mid \underline{\gamma})\, P(\underline{\gamma}) \right)$$

- ✓ Con esta estructura es claro que si el objetivo original es describir a Y, entonces el análisis se puede concentrar en el primer factor y las covariables se pueden considerar fijas.

- ✓ Así el objeto de estudio es simplemente,

$$P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)},\ \underline{\phi})\, P(\underline{\phi})$$

El modelo Normal Lineal

- ✓ La versión más popular de este modelo es la de Regresión Lineal Normal.

$$P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)}, \underline{\phi}) = N(\underline{Y} \mid X\underline{\beta}, \tau I)$$

$$\underline{\phi}^t = (\underline{\beta}, \tau)$$

$$P(\underline{Y}_{(n\times 1)} \mid X_{(n\times k)}, \underline{\phi}) = (2\pi/\tau)^{-n/2} \exp\left[-(\tau/2)(\underline{Y} - X\underline{\beta})^t (\tau I)(\underline{Y} - X\underline{\beta})\right]$$

Normal Multivariada en $\Re^n$

Análisis Conjugado

- ✓ Para efectos del análisis Bayesiano, una posibilidad es utilizar una inicial *conjugada*.

$$P(\underline{\phi}) = P(\underline{\beta}, \tau)$$

$$= P(\underline{\beta} \mid \tau)\, P(\tau)$$

con

$$P(\underline{\beta} \mid \tau) = N(\underline{\beta} \mid \underline{b}, \tau Q)$$

$$P(\tau) = \text{Gamma}(\tau \mid \alpha, \nu)$$

Análisis Conjugado

$$P(\underline{\beta} \mid \tau) = N(\underline{\beta} \mid \underline{b}, \tau Q)$$

$$P(\tau) = \text{Gamma}(\tau \mid \alpha, \nu)$$

Normal Multivariada en $\Re^k$

Q matriz de precisión

$$P(\underline{\beta}, \tau) = P(\underline{\beta} \mid \tau)\, P(\tau)$$

$$= \text{Normal}(\underline{\beta} \mid \underline{b}, \tau Q)\ \text{Gamma}(\tau \mid \alpha, \nu)$$

$P(\underline{\beta}, \tau)$ es una Normal-Gamma $(\underline{\beta}, \tau \mid \underline{b}, Q, \alpha, \nu)$

Análisis Conjugado

- La Normal-Gamma es conjugada para este modelo.

✓ $P(\underline{\beta}, \tau \mid \underline{Y}_{(n\times 1)}\, X_{(n\times k)}) = \text{Normal-Gamma}(\underline{\beta}, \tau \mid \underline{b}_D, Q_D, \alpha_D, \nu_D)$

$$(\underline{b}, Q, \alpha, \nu) \rightarrow (\underline{b}_D, Q_D, \alpha_D, \nu_D)$$

Análisis Conjugado

- Si $X_{(n\times k)}$ es de rango completo $k \leq n$, entonces X^tX es $k\times k$ de rango completo y existe

$$\underline{\hat{\beta}} = (X^tX)^{-1} X^t \underline{Y} \quad , \quad \text{el estimador de MCO de } \underline{\beta}.$$

- También se puede calcular la Suma de Cuadrados de los Errores

$$SCE = (Y - X\underline{\hat{\beta}})^t (Y - X\underline{\hat{\beta}})$$

Análisis Conjugado

- Con estos elementos,

$$(\underline{b}, Q, \alpha, \nu) \rightarrow (\underline{b}_D, Q_D, \alpha_D, \nu_D)$$

- Donde

  - ✓ $Q_D = (X^t X + Q)$
  - ✓ $\underline{b}_D = Q^{-1} ((X^t X)\hat{\underline{\beta}} + Q\underline{b})$
  - ✓ $\alpha_D = \alpha + n / 2$
  - ✓ $\nu_D = \nu + \frac{1}{2} (\underline{b} - \underline{b}_D)^t Q\underline{b} + \frac{1}{2} (\underline{Y} + X\underline{b}_D)^t \underline{Y}$

Análisis Conjugado

- En particular, $P(\tau \mid \underline{Y}_{(n\times 1)}, X_{(n\times k)}) = \text{Gamma}(\tau \mid \alpha_D, \nu_D)$

  con

$$\alpha_D = \alpha + n/2$$

$$\nu_D = \nu + \frac{1}{2}(\underline{b} - \underline{b}_D)^t Q\underline{b} + \frac{1}{2}(\underline{Y} + X\underline{b}_D)^t \underline{Y}$$

- También, $P(\underline{\beta} \mid \underline{Y}_{(n\times 1)}, X_{(n\times k)}) = \text{Student}(\underline{\beta} \mid \underline{b}_D, Q_D\alpha_D/\nu_D, 2\alpha_D)$

  con

$$Q_D = (X^tX + Q)$$

$$\underline{b}_D = Q_D^{-1}((X^tX)\hat{\underline{\beta}} + Q\underline{b})$$

Análisis Conjugado

- El tema de pronósticos se puede abordar con relativa facilidad.
- Si interesa producir r pronósticos simultáneos $y_1, y_2, \ldots, y_r$.
- Si y_i se ha de observar en condiciones $\underline{x}_i$; $i = 1, \ldots, r$.

Entonces, si

- ✓ $\underline{y}$ es el vector de los r valores a pronosticar,
- ✓ X es la matriz ($r \times$ k) de valores de las covariables,

Análisis Conjugado

$$P(\underline{y} \mid \mathcal{X}, \underline{\phi}) = N(\underline{y} \mid \mathcal{X}\underline{\beta}, \tau I)$$

Normal Multivariada en $\Re^r$

- Dados $\mathcal{X}$, $\underline{\beta}$ y τ, este modelo también se puede formular como

$$\underline{y} = \mathcal{X}\underline{\beta} + \underline{\varepsilon}$$

donde $\underline{\varepsilon} \sim N(\underline{\varepsilon} \mid \underline{0}, \tau I)$.

Normal Multivariada en $\Re^r$

Análisis Conjugado

- Así que si, a posteriori, se remueve el condicionamiento en $\underline{\beta}$ (y sólo se condiciona en τ), $\underline{y}$ resulta una combinación lineal de vectores Normales independientes y,

$$P(\underline{y} \mid \mathcal{X}, \tau) = N(\underline{y} \mid \mathcal{X}\,\underline{b}_D, \tau\,(\mathcal{X}^t\,(X^tX + Q)^{-1}\mathcal{X} + I)^{-1})$$

a posteriori.

Análisis Conjugado

- Si esta predictiva posterior (condicional en τ) se multiplica por la distribución marginal (a posteriori también) para τ (que es una Gamma) no es difícil verificar que

$$P(\underset{\sim}{y} \mid \mathcal{X}) = Stu\left(\underset{\sim}{y} \mid \mathcal{X}\,\underline{b}_D,\ (I - \mathcal{X}^t(X^tX + \mathcal{X}^t\mathcal{X} + Q)^{-1}\mathcal{X} + I)^{-1}(\alpha_D / \nu_D),\ 2\alpha_D\right)$$

a posteriori.

Iniciales mínimo informativas

- Es interesante observar los resultados que se producen si se consideran distribuciones iniciales mínimo informativas o de referencia, en particular si se utilizan iniciales conjugadas mínimo informativas,

  ✓ $P(\underline{\beta} \mid \underline{Y}_{(n\times 1)}\, X_{(n\times k)}) = \text{Student}(\underline{\beta} \mid \underline{b}_D, Q_D\alpha_D / \nu_D, 2\alpha_D)$

$$Q_D = (X^tX + Q) \rightarrow X^tX \quad (\text{si } Q \rightarrow 0)$$

$$\underline{b}_D = Q_D^{-1}((X^tX)\underline{\hat{\beta}} + Q\underline{b}) \rightarrow \underline{\hat{\beta}} \quad (\text{si } Q, b \rightarrow 0)$$

Iniciales mínimo informativas

- ✓ $P(\tau \mid \underline{Y}_{(n\times 1)}\, X_{(n\times k)}) = \text{Gamma}(\tau \mid \alpha_D, \nu_D)$

$$\alpha_D = \alpha + n/2$$

$$\nu_D = \nu + \frac{1}{2}(\underline{b} - \underline{b}_D)^t Q\underline{b} + \frac{1}{2}(\underline{Y} + X\underline{b}_D)^t \underline{Y}$$

$$\alpha_D \rightarrow \quad n/2 \quad (\text{si } \alpha \rightarrow 0)$$

$$\nu_D \rightarrow \quad SCE/2 \quad (\text{si } \alpha, \nu \rightarrow 0)$$

Iniciales mínimo informativas

- En resumen,

  - ✓ $P\left(\underline{\beta} \mid \underline{Y}_{(n\times 1)}\ X_{(n\times k)}\right) \rightarrow \text{Student}\left(\underline{\beta} \mid \underline{\hat{\beta}}, (X^{t}X)\,\hat{\tau}, n\right)$

  - ✓ $P\left(\tau \mid \underline{Y}_{(n\times 1)}\ X_{(n\times k)}\right) \rightarrow \text{Gamma}\left(\tau \mid n/2, SCE/2\right)$

Iniciales mínimo informativas

- Para la predictiva,

$$P(\underline{y} \mid \mathcal{X}) = \text{Stu}\left(\underline{y} \mid \mathcal{X}\underline{b}_D,\ (I - \mathcal{X}^t(X^tX + \mathcal{X}^t\mathcal{X})^{-1}\mathcal{X} + I)^{-1}(\alpha_D / \nu_D),\ 2\alpha_D\right)$$

$$\rightarrow \quad \text{Stu}\left(\underline{y} \mid \mathcal{X}\widehat{\underline{\beta}},\ (I - \mathcal{X}^t(X^tX + \mathcal{X}^t\mathcal{X})^{-1}\mathcal{X} + I)^{-1}\widehat{\tau},\ n\right)$$

Ilustración

- *Ejemplo (Koop, 2003).*

$$y_i = x_i \beta + \varepsilon_i \qquad i = 1,.., 50$$

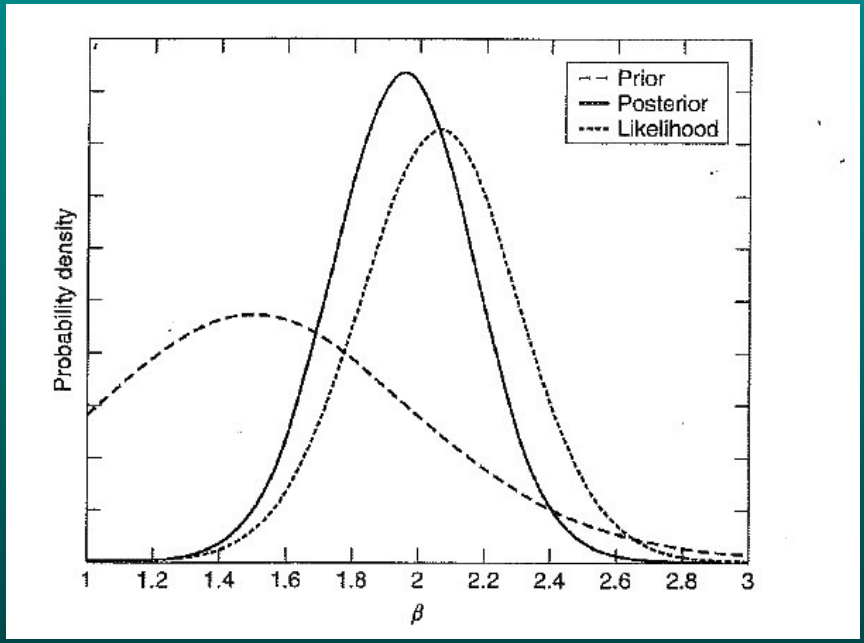
- ✓ $\varepsilon_i \sim N(\varepsilon \mid 0, \tau)$; $i = 1,.., 50$; independientes
- ✓ $\beta = 2, \tau = 1$

$$P(\underline{\beta} \mid \tau) = N(\beta \mid 1.5, \tau\, 0.25)$$

$$P(\tau) = \text{Gamma}(\tau \mid 5, 5)$$

Ilustración

- El caso de β.



Comentarios

- Toda la información que los datos proveen sobre el modelo (combinada con la información inicial) queda capturada en las distribuciones finales

  - ✓ $P(\underline{\beta} \mid \underline{Y}, X)$ para los coeficientes
  - ✓ $P(\tau \mid \underline{Y}, X)$ para la precisión
  - ✓ $P(\underline{y} \mid \underline{x}, \underline{Y}, X)$ predictiva

Comentarios

- Cualquier afirmación sobre la estructura del modelo se puede verificar a partir de las distribuciones finales.

  - $\beta_3 > 0$
  - $\beta_4 = \beta_7$
  - $\beta_6 / \beta_2 < 3$
  - $\tau > 1$
  - $\beta_4 / \tau \in [\ 3, 7\ ]$
  - Si $x = 4 \Rightarrow y > 9$

Comentarios

- Cualquier inferencia se puede plantear como un problema de decisión e involucra una función de pérdida y un mecanismo evaluación (Pérdida Esperada Mínima) que permite discriminar entre inferencias alternativas.

- Existen procedimientos formales para la estimación puntual, la estimación por regiones, la producción de pronósticos y, por supuesto para el contraste de hipótesis.

- El contraste de hipótesis se puede enfrentar como un caso particular del más general problema de selección de modelos.

Comentarios

- La idea general aplica de la misma forma para modelos con estructuras más complejas (heteroscedasticidad y autocorrelación, por ejemplo).

- Lo que ocurre en la práctica es que esas estructuras implican más parámetros en el modelo y una verosimilitud más compleja. En cualquier caso, se debe asignar una distribución inicial conjunta a todos los parámetros desconocidos.

- En muchos de estos casos la distribución inicial también es compleja y, en consecuencia, la posteriori puede presentar dificultades de tratamiento analítico.

Comentarios

- Aun si el modelo se mantiene simple, la especificación de una inicial que puede ser rica en información a priori pero con una estructura no conjugada, puede conducir a una final analíticamente intratable.

- Lo mismo pude ocurrir si una inicial *simple* se combina con una verosimilitud que ocurre en la práctica es que esas estructuras implican más parámetros en el modelo y una verosimilitud más compleja. En cualquier caso, se debe asignar una distribución inicial conjunta a todos los parámetros desconocidos.

Comentarios

- En muchos de estos casos la distribución inicial también que ser mas compleja para reflejar la información disponible sobre el fenómeno y, como consecuencia, la posteriori presenta dificultades de tratamiento analítico.

- Un caso interesante es el de los modelos lineales con *errores Student.* Estos modelos se han utilizado para producir versiones robustas de los modelos Normales que permitan incorporar observaciones con valores atípicos.

Comentarios

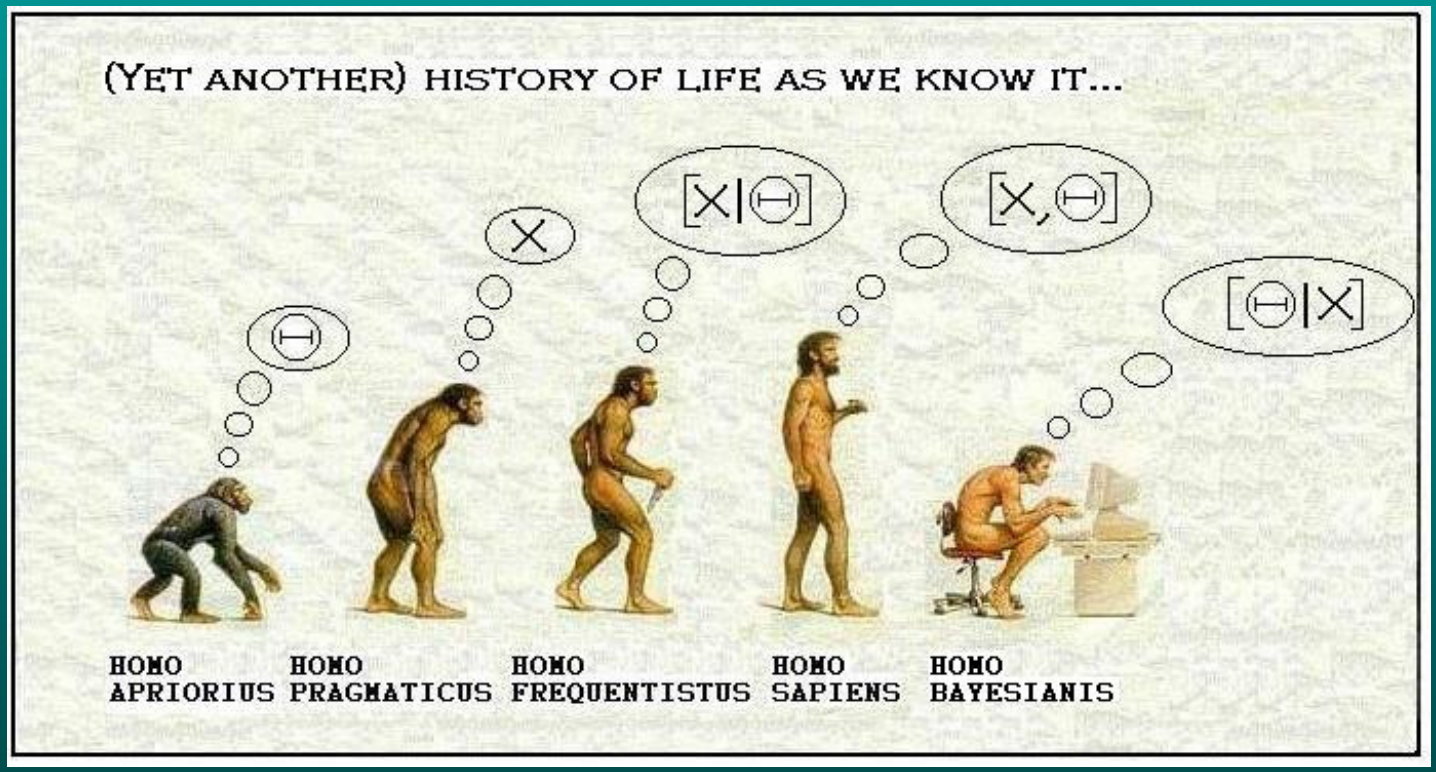
- Cuando se utilizan distribuciones mínimo informativas, la distribución final de los coeficientes preserva la mayor parte de la estructura del caso Normal y la sobre dispersión se captura en la distribución final de la precisión.

- Este tipo de modelos se han generalizado para incluir los casos de distribuciones esféricas y elípticas, en general.

Comentarios

- Los casos de distribuciones intratables analíticamente, se ha resuelto con algoritmos de simulación.
- La idea es simular muestras arbitrariamente grandes de la distribución final de interés y aproximar el cálculo que originalmente debía ser analítico con evaluaciones numéricas a partir de las muestras.
- La familia más popular de algoritmos de este tipo se conoce como MCMC (Markov Chain Monte Carlo) e incluye una variedad mecanismos para simular Cadenas de Markov cuyas distribuciones estacionarias son las distribuciones finales de interés.

Resumen


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS