BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS
POSGRADO EN CIENCIAS MATEMÁTICAS

Ajuste Bayesiano de Curvas

TESIS

QUE PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS (MATEMÁTICAS)

PRESENTA:

Arturo Abraham López Pérez

DIRECTOR DE TESIS

Dra. Gladys Linares Fleites

PUEBLA, PUE., MÉXICO - SEPTIEMBRE 2010

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS
POSGRADO EN CIENCIAS MATEMÁTICAS

Ajuste Bayesiano de Curvas

TESIS

QUE PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS (MATEMÁTICAS)

PRESENTA:

Arturo Abraham López Pérez

DIRECTOR DE TESIS

Dra. Gladys Linares Fleites

PUEBLA, PUE., MÉXICO - SEPTIEMBRE 2010

Índice general

Introducción IX

1. Preliminares 1
1.1. Elementos de la Teoría de la Decisión . . . 1
1.2. Representación Formal . . . 3
1.2.1. Problema de Decisión . . . 3
1.2.2. Principios de Coherencia . . . 5
1.2.3. Principios de Cuantificación . . . 8
1.3. Inferencia Estadística . . . 13

2. Inferencia Bayesiana 15
2.1. Definiciones . . . 15
2.2. El paradigma Bayesiano . . . 17
2.2.1. Intercambiabilidad . . . 18
2.2.2. Teorema de Representación . . . 20
2.3. Inferencia Bayesiana . . . 21
2.3.1. Distribuciones Iniciales . . . 22
2.4. Necesidad de los métodos de simulación . . . 27
2.4.1. Softwares estadísticos Bayesianos . . . 31

3. Ajuste Bayesiano de Curvas 39
3.1. Regresión lineal Bayesiana . . . 40
3.2. Aplicación: Predicción de dióxido de carbono en Veracruz . . . 43
3.2.1. Análisis del Modelo de Regresión Bayesiano . . . 60

A. Enfoque Bayesiano de los procedimientos de la estadística 63
A.1. Naturaleza de la Inferencia Bayesiana . . . 63
A.1.1. Estimación . . . 65
A.1.2. Pruebas de Hipótesis . . . 67

B. Predicción de dióxido de carbono en Veracruz 69
B.1. Enfoque Clásico . . . 69

Dedicado a mi Familia

De pie junto al faro, la luz de la tarde ahuyenta las sombras, el tiempo permanece y el fuego, por fin, toca la vela.

Gracias:

Dra. GLADYS LINARES FLEITES

Por guiarme, pero más por emprender junto a mi el maravilloso laberinto Bayesiano.

Al CONACYT

Por el apoyo económico brindado.

Introducción

La estadística desempeña un importante papel en temas donde interviene la variabilidad, dando lugar a la incertidumbre como parte imprescindible. Pero en la práctica, es común restringir a la estadística a datos, considerandola sólo como el vínculo entre la incertidumbre y los datos. Además, mucha de la literatura en estadística da la impresión de que no existe controversia alguna sobre la validez de los métodos estadísticos en uso, y *ésto no es así*. El campo de la estadística está basado en dos paradigmas fundamentales: Frecuencial y *Bayesiano* [8].

Más allá de los datos, la estadística es esencialmente el estudio de la *incertidumbre*[1], lo que conlleva a la necesidad de investigar un fenómeno desde el enfoque científico. Indudablemente, la estadística no es la única rama del conocimiento que se ha ocupado del estudio de la incertidumbre, la probabilidad examina la manera en como la aleatoriedad en una parte de un sistema afecta a otra, proporcionando a través del modelo de una variable aleotoria o de un proceso estocástico, estimaciones y/o predicciones sobre los datos a producirse, es decir, describe la incertidumbre del fenómeno.

'Medir' la incertudumbre con la probabilidad es, principalmente, compararla con un estándar, de forma que al satisfacer las reglas de la probabilidad sea coherente, además, de que se considera su *interpretación* en las aplicaciones. Éste proceso *pasivo* para describir la incertidumbre a través de la probabilidad, en estadística, es conocido como *inferencia* [24].

El concepto de probabilidad ha ampliado el campo de la estadística más allá de la recopilación y presentación de datos (estadística descriptiva), al tratamiento de datos y al desarrollo de procesos inferenciales (estadística inferencial). Específicamente, la apreciación de la probabilidad que deriva en una formulación basada sobre datos que tienen, por cada valor de un parámetro, una distribución de probabilidad, ha hecho que la estadística se extienda a considerar los datos y las inferencias extraídas de ellos, a la toma decisiones.

La Teoría de la Decisión provee de un paradigma *completo* para la toma de decisiones y la inferencia estadística, proporcionando una base teórica - axiomática que permite construir una metodología estadística general. La metodología estadística frecuencial resulta apropiada en determinados casos y bajo ciertas condiciones, mientras que, la metodología Bayesiana permite abordar de forma sistemática cualquier problema concreto de análisis de datos, sin la necesidad de introducir procedimientos *ad hoc*.

Aunque los problemas específicos pueden ser muy distintos, la existencia de una metodología estadística general permite abordar de forma sistemática cualquier problema concreto de análisis de datos. En un problema estadístico característico (no importando bajo que enfoque sea analizado), existe incertidumbre en los datos y en el parámetro, además, el propósito es profundizar sobre alguna característica del parámetro, a partir de los datos. El enfoque Bayesiano describe la incertudumbre a través de las distribuciones de probabilidad (discretas o continuas) porporcionando, a través de las reglas de la probabilidad[2], un marco para la solución de cualquier problema *inferencial* [26]. La formulación basada sobre datos que tienen una distribución de probabilidad, por cada valor de un parámetro, describe la incertidumbre en los datos y, el concepto de probabilidad sub-

[1] No se intenta aquí discutir el problema de lo que es la incertidumbre, sino, sólo la manera en que puede ser comprendida y manipulada, tal característica.

[2] Específicamente, a través de la regla de *Bayes*.

jetiva permite construir una distribución de probabilidad conjunta sobre todas las características del parámetro[3]- llama *modelo*[4] (de probabilidad).

La modelación sobre la relación existente entre variables es, en particular, un problema que la metodología estadística aborda a través del ajuste de curvas, y entre los métodos paramétricos más extensamente empleados están las *técnicas de regresión.* Éste trabajo pretende describir el estado que guarda la investigación en el tema del análisis del *modelo de regresión lineal Bayesiano*, comenzando por revisar los resultados referentes a la Teoría de la Decisión (Capítulo 1); los principios de coherencia y cuantificación proporcionan el fundamento axiomático de los métodos Bayesianos y garantizan la consistencia de los procedientos a seguir. Posteriormente, se introducen los princípios básicos que se encuentran detrás de la inferencia Bayesiana (Capítulo 2); destacando la importancia del concepto de *Intercambiabilidad*, que a través del *Terorema de representación* integra los conceptos frecuenciales y grantiza la existencia de una distribución inicial para el parámetro (del modelo probabilistico empleado), además, junto al concepto de probabilidad (grado de creencia) determina el proceso inferencial a seguir sobre cualquier magnitud de interés. También, dentro del Capítulo 2, se abordan las diferentes formas para determinar una distribución incial- informativa o no informativa- y, la necesidad de emplear técnicas de simulación para aproximar la distribución final, solución Bayesiana a un problema de inferencia. En particular, se emplean las técnicas Monte Carlo basadas en cadenas de Markov (MCMC) a través de diferentes softwares estadísticos Bayesianos (WinBUGS y R). Una vez establecidos los fundamentos y el proceso inferencial a seguir dentro del paradigma Bayesiano, se aborda de manera sistemática el modelo de regresión lineal Bayesiano (Capítulo 3). Finalmente, se presenta una aplicación sobre la predicción de dióxido de carbono, problema de interés actual por su enlace con los estuidos realizados en México sobre el cambio climático.

En suma, el propósito de este trabajo de tesis, es proporcionar una introducción a la metodología estadística Bayesiana concebida como un instrumento para *analizar* adecuadamente la información de que se dispone y *decidir*, de manera razonable, sobre la mejor forma de actuar.

Arturo Abraham López Pérez
Facultad de Ciencias Físico-Matemáticas
Benemérita Universidad Autónoma de Puebla
Septiembre de 2010

[3]No se esta afirmando que el parámetro sea aleatorio, los axiomas sobre los que se basa la probabilidad subjetiva, permiten deducir que tiene la propiedad de aleatoriedad.

[4]Ésto difiere del concepto comunmente empleado en la estadística frecuencial.

Capítulo 1

Preliminares

El presente capítulo plantea y analiza el problema de decisión. El proceso lógico a seguir para su resolución, proporcionará el *fundamento* sobre el que descansa el concepto de probabilidad, los métodos de inferencia y los criterios de decisión globalmente conocidos como *métodos Bayesianos*.

1.1. Elementos de la Teoría de la Decisión

El estudio de un problema de decisión[1], necesariamente, debe comenzar por considerar las posibles formas de actuar, así, para empezar a solucionar un problema de decisión, hay que elaborar el conjunto de posibles alternativas -*espacio de decisiones*. La construcción de dicho espacio, requiere especial atención, ya que el modelo a estudiar se limitará a elegir uno y sólo uno de sus elementos, por lo que, debe elaborarse un espacio de decisiones *exhaustivo*, que agote todas las posibilidades que puedan parecer razonables. Así mismo, es conveniente exigir que el espacio de decisiones sea de tal forma que la elección de uno de sus elementos *excluya* la elección de cualquier otro, sin suponer pérdida de generalidad.[2]

Determinar la mejor de un conjunto de alternativas sería, en principio, inmediato si se dispone de información completa sobre la consecuencia a que da lugar cada alternativa. Por supuesto, existen situaciones con información completa y, a pesar de ello, es difícil determinar la decisión más adecuada, sin embargo, tal dificultad es técnica, más no conceptual. No será de interés considerar tales dificultades técnicas[3], por ahora, sólo se analiza intuitivamente el *proceso lógico de decisión* en ambiente de incertidumbre, es decir, del método a seguir para la toma de decisiones cuando *no* se dispone de toda la información que se considera relevante, por lo que, el problema general de decisión a considerar, se plantea en un ambiente de incertidumbre.

Determinado el espacio de decisiones, debe tenerse en cuenta para cada una de las decisiones posibles, el conjunto de *sucesos inciertos* que determinan sus eventuales *consecuencias*. En otra palabras, el modelo a considerar exige la especificación de un espacio de decisiones y un conjunto de sucesos inciertos mutuamente excluyentes, los cuales condicionarán las consecuencias de cada decisión. Por lo tanto, un problema de decisión consistirá entonces, en elegir una decisión sin saber cual de los sucesos inciertos tendrá lugar. Si se denota con $D = \{d_1, d_2, ..., d_k\}$, al espacio de decisiones y con ${\Theta_i}^4 = \{\theta_{i1}, \theta_{i2}, ..., \theta_{im_i}\}$ para $1 \leq i \leq k$, al conjunto de los m_i *sucesos inciertos*, esquemáticamente, un problema de decisión puede ser representado mediante un *un árbol de decisión*, del tal forma que *si se opta por la decisión d_i y ocurre el suceso θ_{ij} se obtiene entonces, la consecuencia c_{ij}*.

[1] Problema de elección razonada entre dos o más formas de actuar.

[2] El espacio de decisiones puede ser infinito pero en la mayor parte de las aplicaciones es finito.

[3] Se supondrá que en presencia de información completa siempre es posible elegir la mejor alternativa.

[4] Frecuentemente el conjunto de sucesos inciertos es el mismo cualesquiera sea la decisión por la que se opte, es

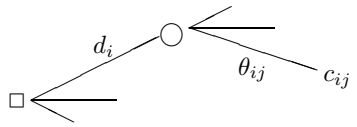


Figura 1.1: Árbol de Decisión

Los sucesos que componen cada Θ_i son inciertos, en el sentido de que no se sabe cual de ellos tendrá lugar y no resultan, en general, igualmente verosímiles. Aunque no se disponga de información suficiente para determinar cual de ellos tendrá lugar, generalmente, la información disponible hace a unos elementos más verosímiles que otros, por lo que, será necesario precisar de alguna forma cuantitativa el contenido de la información incompleta que se posee y, una forma, es asignando a cada suceso un número que mida la verosimilitud que se le atribuye.

La incertidumbre existente sobre la ocurrencia de *algunas* clases de sucesos se ha venido midiendo, como es conocido, por un número al que se le llama *probabilidad*. Para hacerlo, la *unidad* de probabilidad correspondiente a un suceso *cierto* se *distribuye* entre una clase de sucesos mutuamente excluyentes. Si se consideran *igualmente verosímiles* a n sucesos mutuamente excluyentes, se asigna a cada uno de ellos la probabilidad $\frac{1}{n}$. Otra clase de sucesos, que resulta fácil asignar una probabilidad aunque *no* presenten las propiedades de simetría, es de aquellos sucesos que es posible observar repetidamente en *idénticas* condiciones. Considere una situación ε que puede o no dar lugar a un suceso A, y supónga que tal situación puede repetirse indefinidamente. Es un hecho empírico que, frecuentemente, el cociente $\frac{m}{n}$ (el número de veces m que tiene lugar el suceso A entre el número total n de observaciones) se estabiliza y parece tender a un límite, por lo que, parece razonable pensar que, en *ausencia de otra información relevante*, tal límite medirá la verosimilitud del suceso A en las condiciones descritas por ε.

Pero a pesar de la existencia de simetrías o de datos empíricos sobre frecuencias relativas, que es ciertamente *útil* para precisar cuantitativamente la incertidumbre sobre un determinado suceso, no es de modo alguno *necesaria*.

La mayoría de los sucesos relevantes en un problema de decisión no presentan simetrías ni son repetibles, de forma que un concepto de probabilidad restringido a tales clases de sucesos será, inevitablemente, insuficiente [6].

El concepto de *probabilidad* de un suceso A en una situación H, que se utilizará a lo largo de éste trabajo, será la medida de la verosimilitud que se concede al suceso A bajo la condición descrita por H, esto es, la medida del *grado de creencia* en A que sugiere la información contenida en H[5]. Así, la información que se posee sobre la verosimilitud de los distintos sucesos inciertos en cada Θ_i, puede cuantificarse *distribuyendo* la unidad de probabilidad sobre los sucesos $\{\theta_{i1}, ..., \theta_{im_i}\}$. Dado que que estos sucesos son mutuamente excluyentes- como se ha establecido- entonces, puede describirse la información disponible sobre su verosimilitud (bajo la condición H) mediante un conjunto de números $\{p(\theta_{ij} \mid d_i, H) : j = 1, 2, ..., m_i\}$ de tal forma que:

$$\text{(i) } \forall i \in I : 0 \leq p(\theta_{ij} \mid d_i, H) \leq 1,$$

$$\text{(ii) } \sum_{j=1}^{m_i} p(\theta_{ij} \mid d_i, H) = 1.$$

Considerese ahora, que se toma la decisión d_i (bajo la condición H) y acontese θ_{ij} entonces, se obtene una consecuencia c_{ij}. Obviamente, se puede tener alguna preferencia entre las distintas

decir, $\Theta_i = \{\theta_{i1}, \theta_{i2}, ..., \theta_{im_i}\} = \{\theta_1, ..., \theta_m\} = \Theta$ para toda $i = \overline{1, k}$.

[5]Debe advertirse que la probabilidad que se le asigna a un suceso es siempre *condicional* a la información que se posee sobre él: no existen probabilidades *absolutas* [6].

consecuencias. En principio, tal preferencia puede ser cuantificada asignando, a cada una de las consecuencias, un número $u(c_{ij})$ que mida su *utilidad*. Por lo tanto, para toda d_i, $i = \overline{1,k}$, se obtenie una utilidad $u(c_{ij})$ con probabilidad $p(\theta_{ij} \mid d_i, H)$.

Una vez especificadas las probabilidades que describen la verosimilitud de los sucesos inciertos y las utilidades que describen las preferencias entre las posibles consecuencias, el problema de decisión tiene ya una solución:

Si la utilidad *media* de la decisión d_i -*utilidad esperada* de d_i- se define como:

$$u^*(d_i) := \sum_{j=1}^{m_i} u(c_{ij}) p(\theta_{ij} \mid d_i, H) \tag{1.1.1}$$

entonces, resultará natural elegir como decisión más razonable, aquella que *maximice la utilidad esperada* entre las k alternativas.[6]

En suma, existe una única forma razonable de tomar decisiones. En primer lugar, es necesario determinar el conjunto de las decisiones posibles y el de aquellos sucesos cuya ocurrencia pueda modificar las consecuencias de la decisión seleccionada. En segundo lugar debe cuantificarse, mediante probabilidades, la verosimilitud asociada a la ocurrencia de tales sucesos. En tercer lugar, se tiene que describir detalladamente las posibles consecuencias a que dan lugar cada una de las decisiones y la preferencia, entre consecuencias, debe ser evaluada y cuantificada en términos de una magnitud común que recibirá el nombre de *utilidad*. Finalmente, considerar aquella decisión que, con base a las probabilidades calculadas, porporcione la máxima utilidad esperada [5, 12].

1.2. Representación Formal

El presente apartado formaliza y expone la notación de la teoría de la decisión siendo ésta la misma que la empleada en [14] donde, además, se pueden consultar las pruebas de los resultados que se presentan a continuación. El proceso lógico a seguir, para optar por una decisión, no intenta describir la forma en que comúnmente se toman las decisiones, no es de interés presentar una teoría descriptiva, sino, una *teoría normativa*.

1.2.1. Problema de Decisión

Un problema general de decisión, es un problema de elección razonada entre un determinado conjunto de alternativas en presencia de incertidumbre sobre algunos de los factores que condicionan las consecuencias de tal elección. Así, la estructura de una situación en la cual las decisiones son elegidas entre cursos alternativos de acciones con consecuencias inciertas, está determinada por tres elementos básicos:

Definición 1.2.1. Una clase de posibles acciones, $\mathcal{A} = \{a_i : i \in I\}$.

Definición 1.2.2. Para cada $a \in \mathcal{A}$, un conjunto de eventos inciertos con estructura de álgebra, $\Theta_a = \{\theta_{aj} : j \in J\}$[7].

Definición 1.2.3. Un conjunto de consecuencias, $\mathcal{C}_a = \{c(a, \theta) : \theta \in \Theta_a\}$.

El modelo de un problema de decisión a considerar exige, pues, la especificación de un espacio $\mathcal{A}$ de acciones y la determinación de los conjuntos Θ_a de sucesos inciertos, los cuales condicionarán las consecuencias de cada decisión, así:

Definición 1.2.4. Un problema de decisión queda definido por la estructura: $\{(\Theta_a, \mathcal{C}_a) : a \in \mathcal{A}\}$.[8]

[6] Critério Bayes para la toma de decisiones [6].
[7] Partición exhaustiva de un conjunto total de posibilidades.
[8] Sin pérdida de generalidad, puede ser asumido que las posibles acciones son mutamente excluyentes, en caso contrario se debe trabajar sobre el producto cartesiano apropiado.

Comparabilidad

La percepción del estado de incertidumbre, resultante de la elección de una acción partícular, depende del estado de información disponible en el momento de la elección[9], por lo que, el conjunto Θ_a- partición del conjunto total de posibilidades- puede restringirse con nueva información, es decir, algunos θ_{aj}'s pueden ser más inverosímiles (o más aún, ilógicamente imposibles), mientras que otros más verosímiles. Por lo tanto, *la preferencia por un escenario de incertidumbre*[10], *depende del valoramiento de las consecuencias involucradas y de la actitud adoptada sobre la incertidumbre correspondiente a cada* Θ_a [14]. Claramente, ésto último es subjetivo, varia conforme nueva información es adquirida, modificando las preferencias globales entre varios cursos de acción.

Así, para representar la estructura de un problema de decisión, también se debe definir la idea de preferencia entre acciones; entendiendose ésta relación, en el sentido de que si sólo existiesen dos posibles acciones se eligirá aquella que, condicionada a la información disponible en ese momento, resulte más admisible.

Definición 1.2.5. Sea $\{(\Theta_a, \mathcal{C}_a) : a \in \mathcal{A}\}$ un problema de decisión. Para $a_1, a_2 \in \mathcal{A}$, la relación de preferencia entre acciones se define como:

$$a_1 \preceq a_2 \text{ (a_1 es no preferible a a_2).}$$

La introducción de una relación binaria de preferencia ($\preceq$), no asume que cada par de opciones, $(a_1, a_2) \in \mathcal{A} \times \mathcal{A}$, pueda necesariamente ser relacionada. En caso contrario, es decir, si $a_1 \preceq a_2$, $a_2 \preceq a_1$ o ambas, pueden derivarse otras relaciones binarias, a saber:

Definición 1.2.6. Para cada $a_1, a_2 \in \mathcal{A}$ tal que $a_1 \preceq a_2$ o $a_2 \preceq a_1$ se tiene:

1. $a_1 \sim a_2$ si y sólo si $a_1 \preceq a_2$ y $a_2 \preceq a_1$.
2. $a_1 \prec a_2$ si y sólo si $a_1 \preceq a_2$ y $a_2 \npreceq a_1$.
3. $a_1 \succeq a_2$ si y sólo si $a_2 \preceq a_1$.
4. $a_1 \succ a_2$ si y sólo si $a_2 \prec a_1$.

En general, las consecuencias de adoptar una determinada acción suelen depender de la ocurrencia de determinados sucesos inciertos, es decir, formalmente:

Definición 1.2.7. Sea $\{(\Theta_a, \mathcal{C}_a) : a \in \mathcal{A}\}$ un problema de decisión.

$$\text{Una opción es el conjunto: } \mathcal{L} = \{c(a, \theta_{aj}) \mid \theta_{aj} : j \in J\}.$$

Es decir, una *opción* es una situación en la que para cada $j \in J$ se obtiene la consecuencia $c(a, \theta_{aj})$ si acontece, θ_{aj}. Así, cada elemento $a \in \mathcal{A}$ es una opción y será denotado por:

$$a = \{c_{aj} \mid \theta_{aj} : j \in J\}.$$

La relación de preferencia entre opciones $\preceq$, condicionada sobre un estado inicial de información M_0, puede ser también empleada para definir una relación binaria sobre $\mathcal{C} \times \mathcal{C}$. Bastará notar que cada consecuencia es, obviamente, un caso particular de una opción:

Si Ω es el *suceso cierto*, una consecuencia c equivale a la opción $\{c \mid \Omega\}$ [14].

Definición 1.2.8. Sea $\{(\Theta_a, \mathcal{C}_a) : a \in \mathcal{A}\}$ un problema de decisón. Para cada $c_1, c_2 \in \mathcal{C}_a$ se define:

[9] M_0- denotará el estado que guarda la información en un momento inical.

[10] Se le llama escenario de incertidumbre, a la elección de una acción $a \in \mathcal{A}$ que da lugar a uno y sólo uno de los eventos inciertos $\theta \in \Theta_a$, aconteciendo la correspondiente consecuencia $c(a, \theta) \in \mathcal{C}_a$.

$$c_1 \preceq c_2 \text{ (c_1 es no preferible a c_2) si y sólo si } \{c_1 \mid \Omega\} \preceq \{c_2 \mid \Omega\}.$$

También, la relación de preferencia entre opciones puede ser empleada para definir una relación sobre $\Theta_a \times \Theta_a$. Esta nueva relación captura la noción intuitiva de que un evento puede ser 'más verosímil' que otro.

Definición 1.2.9. Sea $\{(\Theta_a, \mathcal{C}_a) : a \in \mathcal{A}\}$ un problema de decisión. Para cada $\theta_{ai}, \theta_{aj} \in \Theta_a$ con $i \neq j$, se define:
$\theta_{ai} \preceq \theta_{aj}$ (θ_{aj} es más verosímil que θ_{ai}) si y sólo si para cada $c_1, c_2 \in \mathcal{C}_a$:

$$c_1 \preceq c_2 \text{ entonces, } \{c_2 \mid \theta_{ai}, c_1 \mid \theta_{ai}^c\} \preceq \{c_2 \mid \theta_{aj}, c_1 \mid \theta_{aj}^c\} \text{ [11]}.$$

Intuitivamente, el contenido de la definición (1.2.9), especifíca que al comparar dos opciones dicotómicas que involucran el mismo par de consecuencias y que difieren sólo en los eventos inciertos que involucran, se debe eligir la opción que sea más verosímil que la consecuencia preferida a obtener[12], es decir, $\{c_2 \mid \theta_{ai}, c_1 \mid \theta_{ai}^c\} \preceq \{c_2 \mid \theta_{aj}, c_1 \mid \theta_{aj}^c\}$ si, y sólo si θ_{aj} es más verosímil que θ_{ai}.

La preferencia para opciones, así como para consecuencias y eventos, se ha definido condicionada sobre un estado de información inicial M_0. Si, subsecuentemente, es necesario tomar en cuenta más información, por la consideración del acontecimiento de un posible evento ϑ, entonces, la preferencia entre opciones debe ser descrita por una nueva relación ($\preceq_\vartheta$), la cual debe tomar en cuenta la información inicial y la proveida por ϑ.

La asociación entre las relaciones $\preceq$ y $\preceq_\vartheta$, está dada por:

Definición 1.2.10. Sea $\{(\Theta_a, \mathcal{C}_a) : a \in \mathcal{A}\}$ un problema de decisión y $\emptyset \preceq \vartheta$.

1. $a_1 \preceq_\vartheta a_2 \Longleftrightarrow \forall a \in \mathcal{A} : \{a_1 \mid \vartheta, a \mid \vartheta^c\} \preceq \{a_2 \mid \vartheta, a \mid \vartheta^c\}$.
2. $\theta_i \preceq_\vartheta \theta_j \Longleftrightarrow c_1 \preceq_\vartheta c_2 : \{c_2 \mid \theta_i, c_1 \mid \theta_i^c\} \preceq_\vartheta \{c_2 \mid \theta_j, c_1 \mid \theta_j^c\}$.

En suma, la representación formal de un problema de decisión incluye un álgebra de eventos Θ, un conjunto de consecuencias $\mathcal{C}$, un conjunto de opciones $\mathcal{A}$[13] y una relación binaria $\preceq_\vartheta$[14] sobre $\mathcal{A} \times \mathcal{A}$, por lo que:

Definición 1.2.11. Un problema de decisión se define por: (Θ , $\mathcal{C}$, $\mathcal{A}$, $\preceq$).

1.2.2. Principios de Coherencia

La preferencia entre opciones, formalizada por la relación $\preceq$, provee una base *cualitativa* para comparar opciones y, por extensión, eventos y consecuencias. Los axiomas de *coherencia* (1.2.1, 1.2.2 y 1.2.3), proporcionarán un conjunto minimal de reglas para asegurar que tal comparación es consistente. El procedimiento a seguir será *perceptivo*, más no descriptivo, es decir, los principios de *comportamiento coherente* sólo permitirán reconocer la parte imprescindible al elegir entre alternativas, evitando comenter inconsistencias.

Se asumirá en adelante, que los problemas de decisión son no triviales y que es siempre posible comparar algún par de opciones dicotómicas.

Axioma 1.2.1. Sea (Θ , $\mathcal{C}$, $\mathcal{A}$, $\preceq$) un problema de decisión.

1. $\exists\ c_1, c_2 \in \mathcal{C}$ tal que $c_1 \prec c_2$.

[11] Dado que para toda $c_1 \prec c_2$ se tiene que, $\{c_2 \mid \emptyset, c_1 \mid \Omega\} \prec \{c_2 \mid \Omega c_1 \mid \emptyset\}$ es siempre cierto, es de esperar que $\emptyset \prec \Omega$.

[12] Tal argumento, es aplicable independientemente de la elección de las consecuencias particulares, siempre que las preferencias entre estas últimas se supongan independientes de toda consideración relativa a los acontecimientos $\theta \in \Theta_a$.

[13] El elemento genérico de $\mathcal{A}$ se donotará con $\{c_j \mid \theta_j : j \in J\}$, donde $\{\theta_j : j \in J\}$ es una partición finita del evento seguro Ω, con $\theta_j \in \Theta$ y $c_j \in \mathcal{C}$ para cada $j \in J$.

[14] Relación de preferencia entre opciones dada la ocurrencia del evento $\emptyset \preceq \vartheta$.

2. Para toda $c_i \in \mathcal{C}$ y todo $\theta_i \in \Theta$, es verdadera una de las siguientes preferencias:

$$(\{c_2 \mid \theta_1, c_1 \mid \theta_1^c\} \preceq \{c_2 \mid \theta_2, c_1 \mid \theta_2^c\}) \vee (\{c_2 \mid \theta_2, c_1 \mid \theta_2^c\} \preceq \{c_2 \mid \theta_1, c_2 \mid \theta_1^c\}).$$

La primera condición, de este axioma es natural, afirma que no tiene sentido hablar de un problema de decisión, si todas las consecuencias son equivalentes, ya que cualquier elección resultaría siempre en la misma consecuencia. La segunda condición, simplemente, establece que al menos se puede expresar una preferencia entre opciones dicotómicas, es decir, equivale a suponer que un problema de decisión en ambiente de incertidumbre tiene una solución.

Axioma 1.2.2. Sea $(\Theta\ , \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión.

1. Para toda $a \in \mathcal{A} : a \preceq a$.

2. Sean $a_1, a_2, a_3 \in \mathcal{A}$. Si $a_1 \preceq a_2$ y $a_2 \preceq a_3$ entonces, $a_1 \preceq a_3$.

Este axioma establece la coherencia entre preferencias bajo la relación de orden definida sobre $\mathcal{A} \times \mathcal{A}$.

Las siguientes proposiciones son una consecuencia directa del axioma 1.2.2.

Proposición 1.2.1. *Sea $(\Theta\ , \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión.*

1. *Para toda $\theta \in \Theta : \theta \preceq \theta$.*

2. *Sean $\theta_1, \theta_2, \theta_3 \in \Theta$. Si $\theta_1 \preceq \theta_2$ y $\theta_2 \preceq \theta_3$ entonces, $\theta_1 \preceq \theta_3$.*

Demostración. Se sigue de la definición 1.2.9 y el axioma 1.2.2.

□

Proposición 1.2.2. *Sea $(\Theta, \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión.*

1. *Si $a_1 \sim a_2$ y $a_2 \sim a_3$ entonces, $a_1 \sim a_3$.*
   ($\theta_1 \sim \theta_2$ y $\theta_2 \sim \theta_3$ entonces, $\theta_1 \sim \theta_3$)

2. *Si $a_1 \prec a_2$ y $a_2 \sim a_3$ entonces, $a_1 \prec a_3$.*
   ($\theta_1 \prec \theta_2$ y $\theta_2 \sim \theta_3$ entonces, $\theta_1 \prec \theta_3$)

Demostración. (1) Sean $a_1, a_2, a_3 \in \mathcal{A}$ tal que $a_1 \sim a_2$ y $a_2 \sim a_3$, de la definición 1.2.8 se tiene, $a_1 \preceq a_2$, $a_2 \preceq a_1$ y $a_2 \preceq a_3$, $a_3 \preceq a_2$. Entonces, por el axioma 1.2.2, $a_1 \preceq a_3$ y $a_3 \preceq a_1$. Por lo tanto $a_1 \sim a_3$[15].

□

El siguiente axioma, formaliza la idea de que la preferencia entre consecuencias es no afectada por la obtención de nueva información.

Axioma 1.2.3. Sea $(\Theta\ , \mathcal{C}, \mathcal{A}, \preceq)$ y $\emptyset \prec \vartheta$ un evento cualesquiera.

1. Si $c_1 \preceq c_2$ entonces, $c_1 \preceq_\vartheta c_2$.

2. Si $c_1 \prec c_2$ tal que $\{c_2 \mid \theta_1, c_1 \mid \theta_1^c\} \preceq \{c_2 \mid \theta_2, c_1 \mid \theta_2^c\}$ entonces, $\theta_1 \preceq \theta_2$.

3. Si para algún c y $\emptyset \prec \vartheta$ tal que $\{a_1 \mid \vartheta, c \mid \vartheta^c\} \preceq \{a_2 \mid \vartheta, c \mid \vartheta^c\}$ entonces, $a_1 \preceq_\vartheta a_2$.

[15] La proposición 1.2.1 y un argumento similar, es empleado para demostrar el caso para eventos.

La condición (2) y (3), aseguran la operacionalidad de la definición 1.2.9 y 1.2.10. En particular (2), afirma que si $\{c_2 \mid \theta_1, c_1 \mid \theta_1^c\} \preceq \{c_2 \mid \theta_2, c_1 \mid \theta_2^c\}$ para algún $c_1 \prec c_2$ entonces, se debe tener ésta preferencia para toda $c_1 \prec c_2$. Esto formaliza la idea intuitiva de que la preferencia sólo debe depender de la 'verosimilitud relativa' de θ_1 y θ_2, más no de las consecuencias empleadas al construir las opciones. Similarmente, (3) afirma que si $\{a_1 \mid \vartheta, c \mid \vartheta^c\} \preceq \{a_2 \mid \vartheta, c \mid \vartheta^c\}$ para algún c entonces, dado ϑ, a_1 no debe preferirse a a_2, así, para cualesquiera $a \in \mathcal{A}$, $\{a_1 \mid \vartheta, a \mid \vartheta^c\} \preceq \{a_2 \mid \vartheta, a \mid \vartheta^c\}$. Este último argumento, es una versión de lo que es llamado:

'Sure Thing Pinciple': Si $(a_1 \preceq a_2 \mid \vartheta) \wedge (a_1 \preceq a_2 \mid \vartheta^c)$ entonces, $a_1 \preceq a_2$ [32, 33].

Una consecuencia importante del axioma 1.2.3, es el hecho de que la preferencia entre consecuencias es invariante bajo cambios en la información respecto a los eventos en Θ.

Proposición 1.2.3. *Sea (Θ, $\mathcal{C}$, $\mathcal{A}$, $\preceq$) un problema de decisión.*

$$c_1 \preceq c_2 \text{ si y sólo si existe } \emptyset \prec \vartheta \text{ tal que } c_1 \preceq_\vartheta c_2.$$

Demostración. ($\Rightarrow$) Si $c_1 \preceq c_2$ entonces, por el axioma 1.2.3-(1), $c_1 \preceq_\vartheta c_2$ para algún ϑ.

($\Leftarrow$)Por la definición 1.2.10-(1), para algún $\emptyset \preceq \vartheta$, $c_1 \preceq_\vartheta c_2$ implica que para alguna opción a se tiene $\{c_1 \mid \vartheta, a \mid \vartheta^c\} \preceq \{c_2 \mid \vartheta, a \mid \vartheta^c\}$. Tomando $a = \{c_1 \mid \vartheta, c_2 \mid \vartheta^c\}$ se sigue que $\{c_1 \mid \vartheta, c_2 \mid \vartheta^c\} \preceq \{c_1 \mid \emptyset, c_2 \mid \Omega\}$. Si $c_1 \succ c_2$, por el axioma 1.2.3-(2), se tiene que $\vartheta \prec \emptyset$ lo cual es una contradicción ya que $\emptyset \prec \vartheta$. Así, por el axioma 1.2.1-(2) se sigue que $c_1 \preceq c_2$.

□

Otra consecuencia importante, del axioma 1.2.3, es sobre el orden de incertidumbre entre eventos respecto a una implicación lógica. Si θ_1 implica a θ_2[16] entonces, θ_2 no puede considerarse menos verosímil que θ_1.

Proposición 1.2.4. *Sea (Θ, $\mathcal{C}$, $\mathcal{A}$, $\preceq$) un problema de decisión.*

$$\text{Si } \theta_1 \subseteq \theta_2 \text{ entonces, } \theta_1 \preceq \theta_2 \text{ [17]}.$$

Demostración. Para algún c_1 y $c_2 \in \mathcal{C}$ tal que $c_1 \prec c_2$, defínase:

$$a_1 = \{c_2 \mid \theta_1, c_1 \mid \theta_1^c\} = \{c_1 \mid \theta_2 \setminus \theta_1, \{c_2 \mid \theta_1, c_1 \mid \theta_1^c\} \mid (\theta_2 \setminus \theta_1)^c\},$$
$$a_2 = \{c_2 \mid \theta_2, c_1 \mid \theta_2^c\} = \{c_2 \mid \theta_2 \setminus \theta_1, \{c_2 \mid \theta_1, c_1 \mid \theta_1^c\} \mid (\theta_2 \setminus \theta_1)^c\}.$$

Por el axioma 1.2.3-(3) con $\vartheta = \theta_2 \setminus \theta_1$ se tiene que $a_1 \preceq a_2$. Así, por la definición 1.2.8, se sigue que $\theta_1 \preceq \theta_2$.

□

Generalmente se trabaja con eventos 'significativos', para los cuales el orden es estricto.

Definición 1.2.12. Sea (Θ , $\mathcal{C}$, $\mathcal{A}$, $\preceq$) un problema de decisión.

Un evento θ es *significativo* dado $\emptyset \prec \vartheta$, si $c_1 \prec c_2$ implica que, $c_1 \prec_\vartheta \{c_2 \mid \theta, c_1 \mid \theta^c\} \prec_\vartheta c_2$[18].

Intuitivamente, un evento significativo dado ϑ, es interpretado como *pácticamente posible más no como cierto* [14], dada la información proporcionada por ϑ.

Proposición 1.2.5. *Sea (Θ , $\mathcal{C}$, $\mathcal{A}$, $\preceq$) un problema de decisión.*

Un evento θ es significativo dado $\emptyset \prec \vartheta$ si y sólo si $\emptyset \prec \theta \cap \vartheta \prec \vartheta$[19].

[16] Es decir, si $\theta_1 \subseteq \theta_2$.
[17] Se sigue que para cualquier evento se tiene $\emptyset \preceq \theta \preceq \Omega$.
[18] Si $\vartheta = \Omega$ simplemente se dice que θ es significativo.
[19] En particular θ es significativo si y sólo si $\emptyset \prec \theta \prec \Omega$.

La esencia operacional del hecho de 'aprender de la experiencia', es que las preferencias pueden cambiar al pasar de un estado de información a otro respecto a la ocurrencia de un evento en Θ y, como caso especial, es cuando estas preferencias no cambian. Esto puede ser descrito en términos de los conceptos hasta ahora definidos.

Definición 1.2.13. Sea $(\Theta, \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión.

$\theta_1 \perp \theta_2$ (θ_1 y θ_2 son independientes) si y sólo si para toda c, c_1, c_2 se tiene:

1. $c \bullet \{c_2 \mid \theta_1, c_1 \mid \theta_1^c\}$ implica que, $c \bullet_{\theta_2} \{c_2 \mid \theta_1, c_1 \mid \theta_1^c\}$,
2. $c \bullet \{c_2 \mid \theta_2, c_1 \mid \theta_2^c\}$ implica que, $c \bullet_{\theta_1} \{c_2 \mid \theta_2, c_1 \mid \theta_2^c\}$[20].

1.2.3. Principios de Cuantificación

Los principios de cuantificación permitirán deducir la necesidad de asignar, a la información inicial, una *medida de probabilidad* que la especifique, y proporcionarán una manera para describir las preferencias entre decisiones, a través de una *función de utilidad esperada*, conllevando así, a elegir la decisión que *maximice la utilidad esperada*.

La preferencia entre opciones, formalizada en la sección anterior, provee sólo una base *cualitativa* para comparar opciones. Tal comparación, es inadecuada para ciertos contextos, por lo que es necesario introducir una cuantificación, *una unidad numérica*, que permita 'medir', de alguna manera, determinados entes por comparación (implícita o explícitamente) con determinada escala numérica.

La precisión, a través de los axiomas de cuantificación, se consigue introduciendo una escala numérica dentro de un contexto dotado con una relación (cualitativa) de orden coherente [14].

Axioma 1.2.4. Existe una subálgebra $\Im \subseteq \Theta$ y una función $\mu : \Im \to [0,1]$ tal que, si $\theta, \theta' \in \Im$ entonces:

1. Si $\theta \preceq \theta'$ si y sólo si $\mu(\theta) \leq \mu(\theta')$.
2. Si $\theta \cap \theta' = \emptyset$ entonces, $\mu(\theta \cup \theta') = \mu(\theta) + \mu(\theta')$.
3. Si $\theta \perp \theta'$ entonces, $\mu(\theta \cap \theta') = \mu(\theta)\mu(\theta')$.
4. Si $\theta_1, \theta_2 \in \Theta$ y $\alpha \in [0,1]$ entonces, existe $\theta \in \Im$ tal que:
   a) $\theta_1 \perp \theta$,
   b) $\theta_2 \perp \theta$,
   c) $\mu(\theta) = \alpha$.
5. Sean $\theta_1, \theta_2 \in \Theta$ tal que $\theta_1 \perp \theta$, $\theta_2 \perp \theta$ y $\theta_1 \perp \theta_2$:

$$\text{Si } \theta_1 \sim \theta \text{ entonces, } \theta_1 \sim_{\theta_2} \theta.$$

Definición 1.2.14. Sea $\theta \in \Theta$. Si $\theta \in \Im$, diremos que θ es un evento estándar.[21]

Definición 1.2.15. $\Im$ se le llamará familia estándar de eventos en Θ [22].

Proposición 1.2.6. *Sea* $\{\alpha_1, \alpha_2, ..., \alpha_n\} \subseteq \Re^+$ *una colección finita de números reales.*

$$\textit{Si } \sum_{i=1}^{n} \alpha_i \leq 1 \textit{ entonces, } \exists \{\theta_i \in \Im : \theta_i \cap \theta_j = \emptyset, i \neq j\} \subseteq \Im\textit{: } \mu(\theta_i) = \alpha_i\textit{, para toda } i = \overline{1,n}.$$

[20] Donde $\bullet$ es una de las relaciones $\prec, \succ$ o $\sim$.
[21] Note que $\emptyset$ y Ω son eventos estándar, ya que $\Im$ es un álgebra.
[22] Se entenderá a Θ como el álgebra generada por los eventos relevantes en el problema de decisíón junto con los eventos de $\Im$.

Demostración. Por el axioma 1.2.4-(4) existe $\theta_1 \in \Im$ tal que $\mu(\theta_1) = \alpha_1$. Para $1 < j \leq n$, suponga inductivamente que $\theta_1, ..., \theta_{j-1} \in \Im$ son tal que $\theta_k \cap \theta_s = \emptyset$ para $k \neq s$. Sea $B_j = \bigcup_{k=1}^{j-1} \theta_k$ y definamos $\beta_j = \sum_{k=1}^{j-1} \alpha_k = \mu(B_j)$. Nuevamente por el axioma 1.2.4-(3) y (4), existe un $T_j \in \Im$ tal que $\mu(B_j \cap T_j) = \mu(B_j)(\frac{\alpha_j}{1-\beta_j})$. Sea $\theta_j = T_j \cap B_j^c$ luego, $\theta_j \cap \theta_i = \emptyset$ para toda $i = \overline{1, j-1}$. Sea $T_j = \theta_j \cup (T_j \cap B_j)$, del axioma 1.2.4-(2) se sigue que, $\mu(T_j) = \mu(\theta_j) + \mu(T_j \cap B_j)$. Por lo que $\mu(\theta_j) = \frac{\alpha_j}{(1-\beta_j)} - \frac{\alpha_j \beta_j}{(1-\beta_j)} = \alpha_j$.

□

Axioma 1.2.5. Sea $(\Theta\ , \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión.

1. Si $c_1 \preceq c \preceq c_2$ entonces, existe $\theta \in \Im$ tal que $c \sim \{c_2 \mid \theta, c_1 \mid \theta^c\}$.
2. Para cada $\theta \in \Theta$, existe $\theta^{'} \in \Im$ tal que $\theta \sim \theta^{'}$.

Este último axioma, reduce a lo más esencial la idea introducida al incio de esta sección, la precisión a través de la cuantificación.

Creencia y Probabilidad

La preferencia entre opciones debe depender, al menos, 'del grado de creencia' sobre los eventos inciertos. Los principios de *coherencia* y *cuantificación* permien dar una definición formal al concepto de *grado de creencia* y proveen de una medida numérica para la incertidumbre asociada a cada evento.

A continuación, se presentan algunos resultados respecto a la relación de incertidumbre entre eventos.

Proposición 1.2.7. *Sea $(\Theta\ , \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión. Si $\theta_1, \theta_2 \in \Theta$ entonces, es cierta una y sólo una de las siguientes afirmaciones:*

$$(\theta_1 \preceq \theta_2) \vee (\theta_1 \sim \theta_2) \vee (\theta_2 \preceq \theta_1).$$

Aunque la relación de orden para opciones, $\preceq$, se asume que es no completa (es decir, no todo par de opciones es comparable), se sigue, como una consecuencia del axioma 1.2.5, que la relación de incertidumbre inducida para eventos es completa (Proposición 1.2.7).

Proposición 1.2.8. *Sea $\theta_i \in \Theta$ para $i = \overline{1, 4}$.*

$$\textit{Si } \theta_1 \preceq \theta_2,\ \theta_3 \preceq \theta_4 \textit{ y } \theta_1 \cap \theta_2 = \emptyset = \theta_3 \cap \theta_4 \textit{ entonces, } \theta_1 \cup \theta_3 \preceq \theta_2 \cup \theta_4.$$

Definición 1.2.16. Dada una relación de incertidumbre $\preceq$.

Se define la *probabilidad* de θ como: $P(\theta) = \mu(\theta^{'})$ con $\theta^{'} \in \Im$ tal que $\theta \sim \theta^{'}$.

Es decir, la probabilidad es un número real asociado a un evento estándar tal que $\theta \sim \theta^{'}$.

Proposición 1.2.9. *Dada una relación de incertidumbre $\preceq$, existe una única probabilidad asociada a cada evento en Θ.*

Demostración. La existencia se sigue del axioma 1.2.5-(2). Para la unicidad, si $\theta \sim \theta^{'}$ y $\theta \sim \theta_1^{'}$ entonces, por la proposición 1.2.2-(2), $\theta^{'} \sim \theta_1^{'}$. Así, el resultado se sigue del axioma 1.2.4-(1).

□

Definición 1.2.17. Una función $f : \Theta \to \Re$ será compatible con una relación de orden $\preceq$ sobre $\Theta \times \Theta$, si para toda $\theta, \theta_1 \in \Theta : \theta \preceq \theta_1$ si y sólo si $f(\theta) \leq f(\theta_1)$.

Proposición 1.2.10. *La función de probabilidad es compatible con la relación de incertidumbre.*

La siguiente proposición, de importancia fundamental, establece que el 'grado de creencia' cuantitativo coherente tiene una estructura de una medida de probabilidad aditiva finita sobre Θ. En otras palabras, *la probabilidad es un grado de creencia coherente*[23] [14].

Proposición 1.2.11. *Sea $\preceq$ una relación de incertidumbre entonces:*

1. *$P(\emptyset) = 0$ y $P(\Omega) = 1$.*
2. *Si $\theta \cap \theta_1 = \emptyset$ entonces, $P(\theta \cup \theta_1) = P(\theta) + P(\theta_1)$.*
3. *Si θ es significativo si y sólo si $0 < P(\theta) < 1$.*

Demostración. (1) De la definición 1.2.16 se sigue que $0 \leq P(\theta) \leq 1$. Más aún, por el axioma 1.2.4-(4) existen θ_* y θ^* tal que $\mu(\theta_*) = 0$ y $\mu(\theta^*) = 1$. Por la proposición 1.2.4 se tiene que $\emptyset \preceq \theta_*$ y, por la proposición 1.2.10 $P(\emptyset) = 0$; de manera similar, $\theta^* \preceq \Omega$ implica que $P(\Omega) = 1$.

(2) Si $\theta = \emptyset = \theta_1$, trivialmente se sigue el resultado. Supongamos que $\theta \succ \emptyset \prec \theta_1$ entonces, por la proposición 1.2.8, $\theta \cup \theta_1 = \theta$; así, si $\alpha = P(\theta)$ y $\beta = P(\theta \cup \theta_1)$ se sigue que $\alpha < \beta$ y, por la proposición 1.2.6, existen eventos $\theta^{'}$ y $\theta_1^{'}$ tal que $\theta^{'} \cap \theta_1^{'} = \emptyset$, $P(\theta^{'}) = \alpha$ y $P(\theta_1^{'}) = \beta - \alpha$. Por la proposición 1.2.7, $\theta_1 \succ \theta_1^*$ o $\theta_1 \sim \theta_1^*$ o $\theta_1 \prec \theta_1^*$. Si $\theta_1 \succ \theta_1^*$ entonces, por la proposición 1.2.8, $\theta \cup \theta_1 \succ \theta^* \cup \theta_1^*$ y por tanto $P(\theta \cup \theta_1) > \beta$ lo cual es imposible ya que $P(\theta \cup \theta_1) = \beta$. Análogamente, si $\theta_1 \prec \theta_1^*$ entonces, $\theta \cup \theta_1 \prec \theta^* \cup \theta_1^*$ y $P(\theta \cup \theta_1) < \beta$ lo cual es contradictorio. Luego, $\theta_1 \sim \theta_1^*$ y por lo tanto $P(\theta_1) = \beta - \alpha$, de donde se sigue el resultado.

(3) Por la proposición 1.2.5, si θ es significativo si y sólo si $\emptyset \prec \theta \prec \Omega$. Luego el resultado se sigue de la proposición 1.2.10.

□

Corolario 1.2.1. *(Estructura de aditividad finita del grado de creencia)*

1. *Sea $\{\theta_j\}_{j \in J} \subseteq \Theta$ es una colección de eventos.*

$$\text{Si } \theta_i \cap \theta_k = \emptyset \text{ para toda } i \neq k \text{ entonces, } P(\bigcup_{j \in J} \theta_j) = \sum_{j \in J} P(\theta_j).$$

2. *Para algún evento $\theta \in \Theta$ se tiene $P(\theta^c) = 1 - P(\theta)$.*

Definición 1.2.18. Si $\{\theta_j\}_{j \in J}$ forma un partición finita de Ω con $P(\theta_j) = p_j$ con $j \in J$ entonces, $\{p_j\}_{j \in J}$ se le llamará una distribución de probabilidad sobre la partición.

Proposición 1.2.12.

1. *$P(\cdot)$ es la única medida de probabilidad compatible con la relación de incertidumbre $\preceq$.*
2. *$\theta \perp \theta_1$ si y sólo si $P(\theta \cap \theta_1) = P(\theta)P(\theta_1)$.*
3. *Sea $\emptyset \prec \vartheta$ y la relación de incertidumbre $\preceq_\vartheta$:*
   a) *$\theta_1 \preceq_\vartheta \theta_2$ si y sólo si $\theta_1 \cap \vartheta \preceq \theta_2 \cap \vartheta$.*
   b) *Si existen $c_1 \prec c_2$ tal que $\{c_2 \mid \theta_1, c_1 \mid \theta_1^c\} \preceq_\vartheta \{c_2 \mid \theta_2, c_1 \mid \theta_2^c\}$ entonces, $\theta_1 \preceq_\vartheta \theta_2$.*

Definición 1.2.19. Dada una relación de incertidumbre $\preceq_\vartheta$ con $\emptyset \prec \vartheta$. La probabilidad condicional de un evento θ_1, dada la ocurrencia de ϑ, es un número real tal que $\theta_1 \sim_\vartheta \theta$ con θ un evento estándar e independiente de ϑ, es decir:

$$P(\theta_1 \mid \vartheta) = \mu(\theta) \text{ tal que } \theta_1 \sim_\vartheta \theta \text{ con } \theta \in \Im \text{ independiente de } \vartheta.$$

[23] Es decir, permite que éste concepto, cuando no sólo se considera una incertidumbre, realice comparaciones coherentes.

Proposición 1.2.13. *Para algún $\emptyset \prec \vartheta$:*

1. $P(\theta \mid \vartheta) = \frac{P(\theta \cap \vartheta)}{P(\vartheta)}$.
2. $\theta_1 \preceq_\vartheta \theta_2$ *si y sólo si* $P(\theta_1 \mid \vartheta) \leq P(\theta_2 \mid \vartheta)$.
3. a) $0 = P(\emptyset \mid \vartheta) \leq P(\theta \mid \vartheta) \leq P(\Omega \mid \vartheta) = 1$.
   b) *Si* $\theta_1 \cap \theta_2 \cap \vartheta = \emptyset$ *entonces,* $P(\theta_1 \cup \theta_2 \mid \vartheta) = P(\theta_1 \mid \vartheta) + P(\theta_2 \mid \vartheta)$.
   c) θ *es significativo dado* ϑ *si y sólo si* $0 < P(\theta \mid \vartheta) < 1$.

Corolario 1.2.2. *Para todo $\emptyset \prec \vartheta$:*

1. *Sea* $\{\theta_j \cap \vartheta\}_{j \in J} \subseteq \Theta$ *es una colección de eventos.*

$$\text{Si } \theta_i \cap \theta_k \cap \vartheta = \emptyset \text{ para toda } i \neq k \text{ entonces, } P(\bigcup_{j \in J} \theta_j \mid \vartheta) = \sum_{j \in J} P(\theta_j \mid \vartheta).$$

2. *Para algún evento* $\theta \in \Theta$ *se tiene* $P(\theta^c \mid \vartheta) = 1 - P(\theta \mid \vartheta)$.

Proposición 1.2.14. *Para cualquier $\emptyset \prec \vartheta$:*

1. $P(\cdot \mid \vartheta)$ *es la única medida de probabilidad compatible con la relación de incertidumbre condicional* $\preceq_\vartheta$.
2. *(Teorema de Bayes)*
   Para alguna partición $\{\theta_j\}_{j \in J}$ *de* Ω*:*

$$P(\theta_i \mid \vartheta) = \frac{P(\vartheta \mid \theta_i) P(\theta_i)}{\sum_{j \in J} P(\vartheta \mid \theta_i) P(\theta_i)}.$$

Este último resultado 1.2.14-(2), es una consecuencia matemática simple, ya que la coherencia cuantitativa implica que el grado de creencia sigue las reglas de la probabilidad.

Dado que $\{\theta_j\}_{j \in J}$ forma un partición y $\sum_j P(\theta_j \mid \vartheta) = 1$[24] entonces, el teorema de Bayes puede ser escrito, para toda $j \in J$, como:

$$P(\theta_j \mid \vartheta) \propto {}^{25} P(\vartheta \mid \theta_j) P(\theta_j),$$

dado que siempre es posible normalizar el producto al dividirlo por la constante de proporcionalidad:

$$P[\vartheta] = \sum_j P(\vartheta \mid \theta_j) P(\theta_j).$$

El teorema de Bayes caracteríza la forma en la cual la creencia inicial, $P(\theta_j)$, sobre los $\theta_j's$ es modificada por nueva información, ϑ, reexaminando y modificando tal creencia por $P(\theta_j \mid \vartheta)$. Este proceso depende esencialmente de la especificación de $P(\vartheta \mid \theta_j)$, la cual refleja como la creencia sobre la nueva información obtenida, varía respecto a los diferentes $\theta_j's$[26][14].

[24] Esta igualdad se sigue del corolario 1.2.2.
[25] $\propto$- denotará *proporcional a*.
[26]

- $P(\theta_j)$ será citada como la *probabilidad a priori* de θ_j, $j \in J$.
- $P(\vartheta \mid \theta_j)$- la *verosimilitud* de ϑ dado θ_j, $j \in J$.
- $P(\theta_j \mid \vartheta)$- la *probabilidad a posteriori* de θ_j, $j \in J$.
- $P(\vartheta) = \sum_{j \in J} P(\vartheta \mid \theta_j) P(\theta_j)$- la *probabilidad predictiva* de ϑ.

Utilidad

La preferencia entre opciones depende, en parte, del 'grado de creencia' asociado a los eventos inciertos involucrados en las opciones pero, también, debe depender del valor asociado a las consecuencias obtenidas. El axioma 1.2.5-(1) provee, de manera intuitiva, una forma de introducir una *medida numérica para las consecuencias* la cual es, al menos, una base operacional coherente [14].

Definición 1.2.20. La consecuencia c_* será la *peor* y c^* será la *mejor* consecuencia en un problema de decisión si, para algún otra consecuencia $c \in \mathcal{C}$:

$$c_* \preceq c \preceq c^*{}^{27}.$$

En adelante, se asumirá que las consecuencias extremas $c_* \prec c^*$ existen en $(\Theta, \mathcal{C}, \mathcal{A}, \preceq)$ *-problema de decisión acotado.*

Definición 1.2.21. Dada una relación de preferencia $\preceq$. Se define la *utilidad* de una consecuencia c relativa a $c_* \prec c^*$ como:

$$u(c) = u(c \mid c_*, c^*) = \mu(\theta), \text{ con } \theta \in \Im \text{ tal que } c \sim \{c^* \mid \theta, c_* \mid \theta^c\}.$$

La función $u : \mathcal{C} \to \Re$ se le llama *Función de Utilidad.*

Proposición 1.2.15. *Si $(\Theta, \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión acotado con consecuencias extremas $c_* \prec c^*$ entonces:*

1. $\forall c \in \mathcal{C}, \exists! u(c \mid c_*, c^*)$.
2. *Si $\theta \in \Theta \setminus \{\emptyset\}$ entonces $u(c \mid c_*, c^*, \theta) = u(c \mid c_*, c^*)$.*
3. $0 = u(c_* \mid c_*, c^*) \leq u(c \mid c_*, c^*) \leq u(c^* \mid c_*, c^*) = 1$.

Los principios de coherencia y cuantificación, han permitido introducir una medida númerica para comparar cuantitativamente eventos y consecuencias, estos principios, también, permiten definir una medida numérica para comparar opciones.

Definición 1.2.22. Sea $(\Theta, \mathcal{C}, \mathcal{A}, \preceq)$ un problema de decisión acotado con consecuencias extremas $c_* \prec c^*$, $\emptyset \prec \vartheta$ y $a = \{c_j \mid \theta_j : j \in J\}$. La utilidad esperada para la opción a dado ϑ respecto a $c_* \prec c^*$ se define como:

$$\bar{u} = (a \mid c_*, c^*, \vartheta) = \sum_{j \in J} u(c_j \mid c_*, c^*) P(\theta_j \mid \vartheta)^{28}.$$

Si $\vartheta = \Omega$, simplemente se escribirá $\bar{u} = (a \mid c_*, c^*)$.

Proposición 1.2.16. *(Criterio de decisión para un problema de decisión acotado)*
Para algún problema de decisión acotado con consecuencias extremas $c_* \prec c^*$ y $\emptyset \prec \vartheta$:

$$a_1 \preceq_\vartheta a_2 \text{ si y sólo si } \bar{u}(a_1 \mid c_*, c^*, \vartheta) \leq \bar{u}(a_2 \mid c_*, c^*, \vartheta).$$

El resultado anterior, depende de la existencia de las consecuencias c_* y c^* y la omisión, de tal suposición, hace necesario seleccionar un par de consecuencias- c_1 y c_2- como referencia, donde no necesariamente $c_1 \preceq c \preceq c_2$ para toda $c \in \mathcal{C}$, generalizando así, la definición 1.2.21 y por tanto, la proposición 1.2.16:

Definición 1.2.23. Dada una relación de preferencia $\preceq$ y $\theta \in \Im$. Definimos la utilidad $u(c \mid c_1, c_2)$ de una consecuencia c, relativa a las consecuencias $c_1 \prec c_2$, como:

[27] No necesariamente estas consecuencias deben existir y petenecer a $\mathcal{C}$.

[28] En el contexto de la teoría probabilidad matemática, si el valor de la utilidad de a es considerado como una 'cantidad aleatoria', sujeta a la ocurrencia de un evento particular θ_j entonces, $\bar{u}$ es simplemente *el valor esperado* de la utilidad de a cuando las probabilidades de los eventos son consideradas condicionadas sobre ϑ.

1. Si $c \prec c_1$ y $c \sim \{c_2 \mid \theta, c \mid \theta^c\}$ entonces, $u(c \mid c_1, c_2) = \frac{-\mu(\theta)}{1-\mu(\theta)}$.

2. Si $c_1 \preceq c \preceq c_2$ y $c \sim \{c_2 \mid \theta, c_1 \mid \theta^c\}$ entonces, $u(c \mid c_1, c_2) = \mu(\theta)$.

3. Si $c \succ c_2$ y $c_2 \sim \{c \mid \theta, c_1 \mid \theta^c\}$ entonces, $u(c \mid c_1, c_2) = \frac{1}{\mu(\theta)}$.

Proposición 1.2.17. *Para todo problema de decisión. Sean $c_1 \prec c_2$- consecuencias y $\emptyset \prec \vartheta$ un evento cualesquiera, entonces:*

$$a_1 \preceq_\vartheta a_2 \text{ si y sólo si } \overline{u}(a_1 \mid c_1, c_2, \vartheta) \leq \overline{u}(a_2 \mid c_1, c_2, \vartheta).$$

Hasta aquí, se han presentado diferentes principios que capturan una colección minima de reglas lógicas necesarias para la toma 'racional' de una decisión[29] llevando, básicamente, a las siguientes conclusiones [12]:

1. La preferencia entre consecuencias debe ser medida con una función de *utilidad*

$$u(c) = u(a, \theta),$$

que especifique, sobre alguna escala numérica, la deseabilidad sobre estas.

2. La incertidumbre de los eventos relevantes debe ser evaluada a través de un conjunto de distribuciones de probabilidad:

$$\{(p(\theta \mid H, a), \theta \in \Theta_a) : a \in \mathcal{A}\},$$

que describa su plausibilidad dada la condicion H, bajo la cual la decisión debe ser seleccionada.

3. La deseabilidad de las acciones es establecida por su correspondiente *utilidad esperada*:

$$\overline{U}(a \mid H) = \int_{\Theta_a} U(a, \theta) p(\theta \mid H, a) d\theta,\ a \in \mathcal{A}^{30}.$$

1.3. Inferencia Estadística

A pesar de que se ha discutido que el desarrollo descrito anteriormente (que no es cuestionable cuando una decisión debe ser elegida) es no aplicable a los problemas de la inferencia estadística, ya que aparentemente no se considera ninguna toma de decisión específica es, sin embargo, demostrable [6] que la inferencia estadística sobre θ tiene realmente la estructura matemática de un problema de decisión, donde la clase de alternativas es el espacio:

$$\mathcal{A} = \{p(\theta \mid D) : p(\theta \mid D) > 0, \int_{\Theta} p(\theta \mid D) d\theta = 1\},$$

[29] Aunque la mayor parte de los problemas de decisión llegan a tener una estructura más compleja que la encapsulada en la definición 1.2.11 es, sin embargo, posible demostrar que los principios de coherencia y cuantificación son suficientes para analizar un problema de decisión secuencial complejo [8].

[30] Frecuentemente, es más conveniente trabajar en términos de la *función de pérdida* definida por $L(a, \theta) = Sup_{a \in \mathcal{A}}\{U(a, \theta)\} - U(a, \theta)$, la cual directamente evalúa, como una función de θ, el 'efecto' de seleccionar una opción errónea. La no deseabilidad relativa de una acción $a \in \mathcal{A}$ es entonces, evaluada por su *pérdida esperada* [25]:

$$\overline{L}(a \mid H) = \int_{\Theta_a} L(a, \theta) p(\theta \mid H, a) d\theta,\ a \in \mathcal{A}.$$

es decir, es el espacio de distribuciones de probabilidad condicional sobre θ dado los datos D, y cuya función de utilidad determina la cantidad de información que, cabe esperar, proporcionen los datos sobre θ.

Por lo tanto, la teoría de la decisión no sólo provee una metodología para ocuparse de problemas de decisión bajo incertidumbre, su desarrollo, en combinación con un enfoque operacional de los conceptos básicos, permite visualizar el problema de la modelación estadística como la identificación o selección de una forma particular de representar la creencia sobre las observaciones, por lo que, la base axiomática de la teoría de la decisión, proporciona consolidación a la lógica del enfoque *Bayesiano*.

Así, la teoría Bayesiana *no* sólo proporciona un conjunto de métodos para la resolución de problemas de decisión concretos, sino, que proporciona además, una base teórica sobre la que puede construirse una teoría general unificada que contiene, como casos particulares, aquellas fórmulas estadísticas clásicas cuya utilidad ha sido demostrada por el tiempo [12].

Capítulo 2

Inferencia Bayesiana

El presente capítulo introduce los principios básicos que se encuentran detrás de la Inferencia Bayesiana. La metodología estadística, presentada a lo largo de este trabajo, tiene su fundamento axiomático en la *teoría de la decisión* y como consecuencia importante, es la necesidad de cuantificar la información sobre la verosimilitud de los distintos sucesos inciertos que resultan relevantes en cada problema concreto. Tal cuantificación debe ser realizada, necesariamente, mediante una medida que satisfaga determinadas propiedades.

El concepto de probabilidad, no será más que la medida de la verosimilitud asociada a la ocurrencia de un suceso en determinadas condiciones [6].

Esto último, caracteriza la escuela del pensamiento en que éste trabajo se inscribe y que recibe el nombre de *Bayesiana*.

2.1. Definiciones

Se asumirá que las distribuciones de probabilidad pueden ser descritas a través de sus funciones de densidad de probabilidad, empleando la misma notación que en [12], además, no se hará distinción entre una cantidad aleatoria y el valor particular que ésta pueda tomar: $\mathbf{x}$ será empleada para representar un vector aleatorio *observable* (generalmente datos) y, θ representará un vector aleatorio no observable (generalmente un parámetro); X y Θ representarán, respectivamente, su dominio. $p(\cdot \mid C)$ representá en general una *densidad de probabilidad* bajo las condiciones C, tal que:

$$\bullet p(\cdot \mid C) \geq 0,$$

$$\bullet \int p(\cdot \mid C) = 1.$$

Definición 2.1.1. Un espacio de probabilidad está definido por los elementos $\{\Omega, \mathcal{F}, P\}$ donde $\mathcal{F}$ es una $\sigma-$ álgebra de Ω y P es una medida de probabilidad aditiva completa sobre $\mathcal{F}$.

Definición 2.1.2. Un vector aleatorio sobre un espacio de probabilidad, es una función:

$$\mathbf{x} : \Omega \to X \subseteq \mathbb{R}^k \text{ tal que } \mathbf{x}^{-1}(B) \in \mathcal{F} \text{ para toda } B \in \mathcal{B}.$$

Definición 2.1.3. La función de distribución de una vector aleatorio $\mathbf{x} \ : \Omega \to X \subseteq \mathbb{R}^k$ sobre $\{\Omega, \mathcal{F}, \mathcal{P}\}$, es una función: $F_{\mathbf{x}} \ : \mathbb{R}^k \to [0,1]$ definida por:

$$F_x(\mathbf{x}) = F_x(x_1, ..., x_k) = P_{\mathbf{x}} \ \{(-\infty, x_1] \cap ... \cap (-\infty, x_k]\} \text{ con } x_i \in \mathbb{R}, 1 \leq i \leq k.$$

Si la distribución de probabilidad se concentra sobre un conjunto numerable de valores, $X = \{x_1, x_2, ...\}$, con $x_i \in \mathbb{R}^k$ para toda $i \in \mathbb{N}$, $\mathbf{x}$ se llama un vector aleatorio *discreto* y el campo escalar $p_x(\mathbf{x}) : \mathbb{R}^k \to [0, 1]$ tal que:

$$p_x(\mathbf{x}) = P\{\omega; \mathbf{x}(\omega) = \mathbf{x}\},$$

es llamado *función de probabilidad*. Si la distribución de probabilidad es tal que existe un campo escalar p_x tal que:

$$P_{\mathbf{x}}(B) = \int_B p_x(\mathbf{x})dt = \int p_{\mathbf{x}}\ (x_1, ..., x_k)dx_1...dx_k, B \in \mathcal{B},$$

entonces $\mathbf{x}$ se llama vector aleatorio *continuo* y, $p_{\mathbf{x}}$ es llamada *función de densidad*.

La tabla 2.1, proporciona diferentes definiciones de funciones de probabilidad y de funciones de densidad a emplear en este trabajo.

Nombre	Densidad de Probabilidad- Función de Masa de Probabilidad	Parámetro(s)
Binomial	$\mathbf{Bi}(x \mid \theta, n) = \binom{n}{x}\theta^x(1-\theta)^{1-x}, x \in \{0, ..., n\}$	$n = 1, 2, ...; 0 < \theta < 1$
Exponencial	$\mathbf{Ex}(x \mid \theta) = \theta e^{-\theta x}, x > 0$	$\theta > 0$
Normal	$\mathbf{N}(x \mid \mu, \sigma) = \sigma^{\frac{1}{2}}(2\pi)^{-\frac{1}{2}} \cdot exp\{-\frac{1}{2}\sigma(x-\mu)^2\}, -\infty < x < \infty$	$-\infty < \mu < \infty, \sigma > 0$
Gamma	$\mathbf{Ga}(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, x > 0$	$\alpha, \beta > 0$
Beta	$\mathbf{Be}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, x \in (0, 1)$	$\alpha, \beta > 0$
ExpGamma	$\mathbf{Eg}(x \mid \alpha, \beta) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}, x > 0$	$\alpha, \beta > 0$
Poisson-Gamma	$\mathbf{Pg}(x \mid \alpha, \beta, n) = \frac{\beta^\alpha\Gamma(\alpha+x)v^x}{\Gamma(\alpha)x!(\beta+v)^{\alpha+x}}, x \in (0, 1, 2, ...)$	$\alpha, \beta, v > 0$
Gamma Invertida	$\mathbf{Ig}(x \mid \alpha, \beta) = \frac{\beta^\alpha x^{-(\alpha+1)}e^{\frac{-\beta}{\alpha}}}{\Gamma(\alpha)}, x > 0$	$\alpha, \beta > 0$
Pareto	$\mathbf{Pa}(x \mid \alpha, \beta) = \alpha\beta^\alpha x^{-(\alpha+1)}, \beta \leq x < \infty$	$\alpha, \beta > 0$

Tabla 2.1: Distribuciones de probabilidad, discretas, continuas, univariadas y multivariadas.

Definición 2.1.4. La esperanza de y, $E[y] = E[g(x)]$, con x y y cantidades aleatorias y $y = g(x)$, se define como:

$$\int_Y yp_y(y)dy = \int_X g(x)p_x(x)dx.$$

Definición 2.1.5. Las cantidades aleatorias $x_1, ..., x_n$, son mutuamente independientes si, para algún $t_i \in \Re$, los eventos $\{\omega : x_i(\omega) \leq t_i\}$ para $i = 1, ..., n$, son mutuamente independientes.

Definición 2.1.6. Para una cantidad aleatoria y, las cantidades aleatorias $x_1, ..., x_n$ son condicionalmente independientes dado y si, para algún $t_i \in \Re$ con $i = 1, ..., n$, los eventos $\{\omega : x_i(\omega) \leq t_i\}$ son condicionalmente independientes dado el evento $\{\omega : y(\omega) \leq y\}$.

Definición 2.1.7. Sean $x_1, x_2, ...$ una secuencia de cantidades aleatorias. Un modelo de probabilidad predictivo es una medida de probabilidad P, que matemáticamente especifica la forma de la distribución conjunta de creencia para algún subconjunto de $x_1, x_2, ...$

2.2. El paradigma Bayesiano

El modelo matemático conveniente para analizar muchos problemas estadísticos contiene un *espacio muestral* X de elementos x con una apropiada estructura de σ-anillo de conjuntos, sobre los cuales se da una familia de medidas de probabilidad indixada por una cantidad θ llamada *parámetro* perteneciente a un *espacio paramétrico* Θ. Los valores x son denominados de varias maneras, *muestra*, *observaciones* o *datos*. θ sólo identifica una medida de probabilidad, además, el modelo incluye a la estadística no-paramétrica, ya que, técnicas especiales son necesarias dentro de este campo por la complejidad del espacio Θ.

Es suficiente también, para muchos problemas, suponer que la medida de probabilidad es una medida σ-finita[1], por lo que puede describirse a través de su *función de densidad*, $p(x \mid \theta)$, con respecto a dicha medida, $\mu(x)$, en el sentido de que:

$$P(A \mid \theta) = \int_A p(x \mid \theta) d\mu(x), \tag{2.2.1}$$

donde A es algún elemento en el σ-anillo y $P(A \mid \theta)$ es la probabilidad de A, determinada por la medida representada por θ[2]. A pesar de estos requisitos, el modelo descrito es empleado en el análisis de todos los problemas estadísticos, y muchas descripciones van más allá, introduciendo un *espacio de decisión* D de elementos d y una *función de pérdida* $L(d, \theta)$ sobre $D \times \Theta$.

El enfoque Bayesiano extiende el modelo básico en una dirección diferente y, supone que Θ tiene, además, de una estructura de σ-anillo, una medida de probabilidad- $p(\theta)$, sobre él. Ésta suposición, es una propiedad definitoria que constituye al argumento Bayesiano y, una solución Bayesiana es una que emplea tal distribución; por lo que es conveniente describir tal medida a través de su función de distribución con respecto a alguna medida σ-finita[3].

Introduciendo la medida de probabilidad singular sobre el espacio paramétrico, una familia de tales medidas puede ser generada, una por cada punto x, a través del Teorema de Bayes:

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)}, \tag{2.2.2}$$

donde:

$$p(x) = \int_\Theta p(x \mid \theta) p(\theta) d\theta. \tag{2.2.3}$$

Así, todas las distribuciones resultantes, $p(\theta \mid x)$, son determinadas por la misma medida. Es usual referirse a $p(\theta)$ como la densidad *inicial (a priori)*; esto es desafortunado, ya que la propiedad de ser inicial se refiere no a una distribución, sino, a una relación entre una distribución y x (o bien X) [4].

Para muchos problemas, el enfoque Bayesiano va más allá, ya que, emplea un espacio de decisión y una *función de utilidad* acotada, $U(d, \theta)$ sobre $D \times \Theta$, que describe el valor de la decisión d cuando

[1] Generalmente la medida de Lebesgue o una medida continua.

[2] Se supondrá que esto siempre es así, por lo que simplemente se escribirá:

$$P(A \mid \theta) = \int_A p(x \mid \theta) d(x).$$

[3] La integral respecto a tal medida será denotada por: $\int ...d\theta$.

[4] $p(\theta)$ describirá la distribución inicial y $p(\theta \mid x)$ la densidad *final -a posteriori-* (de θ) respecto a x.

el valor del parámetro es θ. Un problema de desición es entonces, solucionado por la maximización de la utilidad esperada (final):

$$\int_{\Theta} U(d,\theta)p(\theta \mid x)d\theta. \tag{2.2.4}$$

Por lo tanto, la solución Bayesiana a un problema de inferencia es, simplemente, obtener la distribución (final) dada en la ecuación (2.2.2). Por lo que, los argumentos Bayesianos proveén un método aplicable a una gran variedad de problemas dentro del marco (X, Θ, D) sin la necesidad de introducir procedimientos *ad hoc* [25].

Un método estrechamente relacionado a la coherencia pero mucho más débil y que, además, es expresando en términos del modelo (X, D, Θ), es el empleo de una función de *pérdida*, $L(d,\theta)$[5], junto con la noción de una función de *decisión*, δ, que mapea X sobre D donde $\delta(x)$ es la decisión a tomar si x es observado. La *función de riesgo* para δ es definida como:

$$R(\delta,\theta) := \int_{x} L[\delta(x),\theta]p(x \mid \theta)dx, \tag{2.2.5}$$

donde a δ se llama inadmisible si existe otra función de decisión δ^* con $R(\delta^*,\theta) \leq R(\delta,\theta)$ para toda θ, en caso contrario admisible. Es demostrable que δ es sólo admisible si es una solución Bayesiana para alguna densidad incial $p(\theta)$.[6]

Así, los argumentos de *coherencia* proveén una descripción completa para un problema de *decisión* en términos de $(X, D, \Theta, p(x \mid \theta), p(\theta), U(d,\theta))$, es decir, a través de las leyes de probabilidad, el principio de maximización de la utilidad esperada y el marco formal para la resolución de cualquier situación de decision. Dentro del paradigma Bayesiano, un problema de *inferencia* es descrito en términos de $(X, \Theta, p(\theta), p(x \mid \theta))$ y solucionado por el cálculo de $p(\theta \mid x)$ o alguna marginal, siendo los argumentos de coherencia aplicados directamente a los eventos de Θ [25].

Los resultados expuestos, afirman que cualquier consideración razonable para la toma de decisiones, o inferencias coherentes, lleva a la existencia de $p(\theta)$, $U(d,\theta)$ y al principio de la maximización de la utilidad esperada; particularmente, dado la influencia de estas ideas, los argumentos establecen la existencia de una distribución sobre Θ, a pesar de que se afirma que 'una distribución inicial no existe para algún problema', ésto, dentro del marco de coherencia, es demostrable que es no cierto. Así mismo, la formalización del concepto de probabilidad como *grado de creencia*, deducido a partir de la teoría de la decisión, permite un tratamiento riguroso de tal concepto[7]. Esto no debe ser confundido con lo que se llama probabilidad objetiva que deriva de las consideraciones frecuenciales, la relación entre estas dos ideas es explicable bajo el concepto de *intercabiabilidad* que, a través del *teorema de representación*, permite integar un paradigma unificado de los conceptos estadísticos frecuenciales asociados a modelos paramétricos con tal concepto de probabilidad (grado de creencia).

2.2.1. Intercambiabilidad

El concepto general de *intercambiabilidad* permite modelar, más flexiblemente, muchos experimentos establecidos y, a través del teorema de representación para una secuencia de cantidades aleatorias intercambiables, establece que cualquier análisis coherente de la información requiere la especificación de una distribución conjunta de probabilidad sobre todos los parámetros involucrados, forzando así, a un enfoque *Bayesiano* [9, 11].

[5] A pesar de que es no obvio que tal función deba existir y su significado se un tanto obscuro, en la aplicaciones parece generalmente cierto que: $L(d,\theta) = sup_d U(d,\theta) - U(d,\theta)$, por lo que será conciderado como tal.

[6] $p(\theta)$ puede ser también una incial *impropia*, es decir, una para la cual $\int_{\Theta} p(\theta)d\theta$ diverge.

[7] El punto de vista sobre la probabilidad, que emerge a partir de tal sistema axiomatico, es considerado enteramente subjetivo.

La única descripción satisfactoria de la incertidumbre es por medio de probabilidad, por lo que, la información que se dispone en las condiciones H sobre el valor de una observación futura $x \in X$, debe describirse mediante una densidad de probabilidad $p_x(\cdot \mid H)$, en el sentido de que:

$$P(x \in A \subset X \mid H) = \int_A p_x(x \mid H)dx. \tag{2.2.6}$$

Sea $\{x_1, ..., x_n\}$ una base de datos de n observaciones 'semejantes' a x. Para 'predecir' el valor de x, debe especificarse la correspondiente densidad de probabilidad condicional:

$$p(x \mid x_1, ..., x_n, H).^8 \tag{2.2.7}$$

Si el modelo $p(x_1, ..., x_n \mid H)^9$ especifica, para toda n, la densidad conjunta pero, si se asume que éste puede ser escrito en la forma:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i \mid H), \tag{2.2.8}$$

entonces las x_i's son cantidades aleatorias *independientes*, por lo que, directamente se sigue que para algún $1 \leq m < n$:

$$p(x_{m+1}, ...x_n \mid x_1, ..., x_m, H) = p(x_{m+1}, ...x_n \mid H), \tag{2.2.9}$$

de modo que ningún aprendizaje de experiencia puede ocurrir dentro de esta secuencia de observaciones, es decir, los datos pasados no proveén información adicional sobre los posibles resultados de observaciones futuras en la secuencia. Por lo tanto, el modelo predictivo especifica que tal estructura *independiente* es claramente inapropiada en contextos donde se cree que la acumulación suceciva de datos proporcionará información sobre eventos futuros. En tales casos, la estructura de la densidad conjunta, $p(x_1, ..., x_n \mid H)$, puede encapsular alguna forma de *dependencia* entre las candidades aleatorias $x_1, ..., x_n$.

Si un conjunto $\{x_1, ..., x_n, x_{n+1}\}$ de $n+1$ observaciones, bajo la condición H, es descrito por su densidad de probabilidad conjunta, $p(x_1, ...x_{n+1} \mid H)$, de forma que:

$$P(\{x_1, ..., x_n, x_{n+1}\} \in B \mid H) = \int_B p(x_1, ..., x_n, x_{n+1} \mid H)dx_1...dx_{n+1}, B \subset X^{n+1}, \tag{2.2.10}$$

y consecuentemente:

$$P(x_{n+1} \in A \mid x_1, ..., x_n, H) = \int_A p(x_{n+1} \mid x_1, ..., x_n, H)dx_{n+1}, A \subset X, \tag{2.2.11}$$

con:

$$p(x_{n+1} \mid x_1, ..., x_n, H) = \frac{p(x_1, ..., x_n, x_{n+1}, H)}{p(x_1, ..., x_n, H)},$$

entonces, para predecir el valor de x, es necesario y suficiente especificar una distribución conjunta para cualesquiera sucesión de observaciones 'semejantes', es decir, la densidad conjunta $p(x_1, ..., x_n, H)$, debe encapsular el tipo de *dependencia* asumido entre las cantidades aleatorias x_i[10]

Definición 2.2.1. Sea $\{x_1, ..., x_n, ...\}$ una secuencia de cantidades aleatorias. Diremos que son *intercambiales* bajo una medida de probabilidad P si, para cada subconjunto finito de tal secuencia, su distribución conjunta satisface:

[8] Observe que no se introducen 'parámetros' desconocidos: $p(x \mid x_1, ..., x_n, H)$ es una función totalmente especificada.

[9] Modelo de probabilidad predictivo.

[10] En general, existe un gran número de posibles suposiciones sobre la forma que puede tomar tal dependencia, pero hay algunas formas particularmente simples que describen con exactitud una gran clase de disposiciones experimentales.

$$P(x_1, ..., x_n) = P(x_{\pi(1)}, ..., x_{\pi(n)}),$$

para todas las permutaciones π definidas sobre el subconjunto $\{1, ..., n\}$.

La definición anterior establece que toda información relevante, en un cojunto de variables intercambiales, está contenida en los valores de las $x_i's$, de forma que sus índices son 'no informativos'[11]. Así, el concepto de intercabiabilidad generaliza el de independencia condicional: un conjunto de observaciones independientes idénticamente distribuidas son *siempre* un conjunto de observaciones intercambiables[12] [11].

2.2.2. Teorema de Representación

La suposición de 'similitud' en la intercambiabilidad tiene implicaciones matemáticas importantes, una de ellas, el teorema de representación para variables dicotómicas [9].

Teorema 2.2.1. *Si $\{x_1, x_2, ...\}$, $x_i \in \{0, 1\}$ son variables aleatorias intercambiables entonces:*

- *Existe $\theta \in (0, 1)$ tal que: $\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i$.*
- *Existe $p(\theta)$- densidad de probabilidad.*

Tal que la densidad conjunta, $p(x_1, ..., x_n)$, tiene la representación integral:

$$p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} p(\theta) d\theta.$$

Este teorema, demuestra que un conjunto de variables aleatorias dicotómicas intercambiables *necesariamente* se comporta como una muestra aleaotria de observaciones Bernoulli, cuyo parámetro θ es el límite de su frecuencia relativa, y para el que *necesariamente* existe una distribución inical $p(\theta)$.

En *general*, el teorema de representación, proporciona una percepción integral de la densidad conjunta $p(x_1, ..., x_n)$ para cualquier subconjunto de cantidades aleatorias intercambiables [14].

Teorema 2.2.2. *Si $\{x_i\}_{i\in I}$ es una secuencia infinita de cantidades aleatorias reales intercambiables con medida de probabilidad P entonces, existe una medida de probabilidad Q sobre $\Im$[13] tal que la función de distribución conjunta de $x_1, ..., x_n$ tiene la forma:*

$$P(x_1, ..., x_n) = \int_{\Im} \prod_{i=1}^{n} F(x_i) dQ(F),$$

donde:

$$Q(F) = \lim_{n\to\infty} P(F_n),$$

con F_n- función de distribución empírica definida por $x_1, ..., x_n$.

[11] La información que proporciona cada x_i es independiente de su orden.

[12] Es importante notar que si las observaciones son condicionalmente independientes- hecho asumido implícitamente cuando se considera que las observaciones son una muestra aleatoria de algún modelo- entonces, necesariamente son intercambiables.

[13] Espacio de todas las funciones de distribución sobre $\Re$.

Así, para el caso dicotómico, la intercambiabilidad identifica las observaciones como una muestra aleatoria de un modelo probabilístico específico (Bernoulli) y garantiza la existencia de una distribución y, en general, para vaiables aleatorias de cualquier rango y dimensión, la intercambiabilidad identifica las observaciones como una muestra aleatoria de *algún* modelo probabilístico y garantiza la existencia de una distribución inicial sobre el parámetro que lo describe.

Especificamente, si $\{x_1, ..., x_n, ...\}$ es una secuencia intercambiable de cantidades aleatorias reales entonces, *existe un modelo paramétrico*, $p(x \mid \theta)$, determinado por algún parámetro $\theta \in \Theta$, el cual es el límite (cuando $n \to \infty$) de alguna funcion de los $x_i's$ y, *existe* una distribución de probabilidad para θ, con densidad $p(\theta)$ tal que:

$$p(x_1, ..., x_n) = \int_\Theta \prod_{i=1}^{n} p(x_i \mid \theta) p(\theta) d(\theta), \tag{2.2.12}$$

es decir, si una secuencia de observaciones es considerada intercambiable entonces, para cualquier subconjunto finito de observaciones, pueden ser siempre *interpretadas* como una muestra aleatoria de algún modelo $p(x_i \mid \theta)$, *existiendo* una *distribución inicial* $p(\theta)$, que describe la información inicial sobre el parámetro que determina el modelo [9].

Por lo tanto, el teorema de la representación[14]- resultado de la teoría de probabilidad- demuestra que si las observaciones se consideran intercambiables, pueden de hecho ser una muestra aleatoria, existiendo una distribución de probabilidad inicial sobre el parámetro del modelo.

2.3. Inferencia Bayesiana

El concepto de probabilidad como grado de creencia proporciona, junto al teorema de representación, un paradigma unificado para la inferencia estadística Bayesiana [9].

Sea $\{x_1, ..., x_n\}$ un conjunto de observaciones intercambiables, el proceso de inferencia sobre una magnitud de interés, ϕ, relacionada a tal conjunto se descompone, de forma natural, en tres fases:

1. Especificación de un modelo probabilístico.
2. Especificación de una distribución inicial.
3. Determinación de la distribución final de la magnitud de interés.

- *Especificación del modelo.*
  De la intercambiabilidad se las $x_i's$ se deduce que un conjunto de observaciones, $\{x_1, ..., x_n\}$, constituye una muestra aleatoria de algún modelo probabilístico $p(x \mid \theta)$, de forma que las $x_i's$ son condicionalmente independientes dado el valor del parámetro θ (con $\theta = \lim_{n\to\infty} f(x_1, ..., x_n)$[15]), es decir:

$$p(x_1, ..., x_n \mid \theta)^{16} = \prod_{i=1}^{n} p(x_i \mid \theta), \tag{2.3.1}$$

- *Especificación de la distribución inicial.*
  También, de la intercambiabilidad de las $x_i's$, se deduce la existencia de una distribución inicial $p(\theta)$ sobre el parámetro del modelo probabilístico. Si se dispone de información relevante sobre el valor del θ, esta información debe ser descrita por medio de $p(\theta)$ empleando una

[14] El teorema de representación es sólo un teorema de *existencia*: no especifica el modelo ni la distribución inicial requerida.

[15] Es decir, $\theta \in \Theta$ se define como el límite de alguna función de las observaciones.

[16] Frecuentemente, la información de que se dispone en el contexto en que las observaciones han sido obtenidas, permite especificar el modelo.

distribución inicial subjetiva y, en caso contrario, si no se dispone de información relavante sobre θ entonces, debe emplearse una distribución inicial no informativa, donde la forma de $p(\theta)$, depende necesariamente del contexto..

- *Determinación de la distribución final.*
  La magnitud de interés puede ser tanto una función de los parámetros del modelo, $\phi = \phi(\theta)$, como una función de un conjunto de observaciones futuras, $\phi = \phi(x_{n+1}, ..., x_{n+m})$. En ambos casos, su determinación es un problema bien definido del cálculo de probabilidades.

  Es necesario determinar, en primer lugar, la distribución del parámetro del modelo mediante el teorema de Bayes, es decir, de la teoría de la probabilidad, se sigue que la información disponible sobre el valor del parámetro θ, una vez dado el resultado $\mathbf{x} = \{x_1, ..., x_n\}$, es descrita por la densidad posterior:

$$p(\theta \mid x_1, ..., x_n) = \frac{p(x_1, ..., x_n \mid \theta)p(\theta)}{p(x_1, ..., x_n)} = \frac{\prod_{i=1}^{n} p(x_i \mid \theta)p(\theta)}{p(x_1, ..., x_n)}. \tag{2.3.2}$$

  Si la magnitud de interés es una función de los parámetros del modelo, $\phi = \phi(\theta)$, su distribucion final, $p(\phi \mid x_1, ..., x_n)$, puede deducirse de $p(\theta \mid x_1, ..., x_n)$, como la correspondiente distribución marginal. Si es una función de observaciones futuras, $\phi = \phi(x_{n+1}, ..., x_{n+m})$ entonces, la distribución condicional, $p(\phi \mid \theta)$, puede ser deducida a partir del modelo y, la distribución final de ϕ, puede calcularse utiliando el teorema de la probabilidad total, de forma que:

$$p(\phi \mid x_1, ..., x_n) = \int_{\Theta} p(\phi \mid \theta)p(\theta \mid x_1, ..., x_n)d\theta^{17}. \tag{2.3.3}$$

2.3.1. Distribuciones Iniciales

Las distribuciones iniciales son parte fundamental de la inferencia Bayesiana. Indudablemente, el punto más crítico y criticado del análisis Bayesiano es la elección de la distribución inicial, su determinación, es por tanto lo más importante para realizar inferencias y, hasta cierto punto, es también difícil.

En la inferencia clásica, la muestra de datos es seleccionada aleatoriamente mientras que el parámetro poblacional es considerado fijo, en el análisis Bayesiano, los parámetros siguen una distribución de probabilidad, conocimiento que (antes de considerar los datos) es simplificado en una distribución inicial [18, 19, 20, 41].

Una parte central en la estadística Bayesiana, es el *Teorema de Bayes*, el cual puede reescribirse como:

$$\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta)^{18},$$

donde $p(x \mid \theta)$ es el modelo probabilistico, para un conjunto de observaciones dado, y $\pi(\theta)$ la función inicial. Dadas estas dos funciones es 'fácil' calcular la distribución final de θ, es decir, $\pi(\theta \mid x)$, la cual es empleada para hacer inferencias desde el enfoque Bayesiano. El problema es, pues, el como definir la distribución inicial $\pi(\theta)$. En esta sección describiremos, someramente, la manera en que se puede proceder.

Si se cuenta con información inicial sobre el parámetro θ, esta debe incorporarse en la densidad inicial (*inicial informativa*) o bien, si el conocimiento inicial sobre θ puede ser empleado para especificar una distribución inicial con una forma funcional particular entonces, una familia paramétrica de funciones puede ser definiada (*inicial conjugada*).

[17] Por supuesto que los resultados obtenidos en un análisis como el descrito, son obviamente, condicionales al modelo probabilístico y a la distribución inicial empleada.

[18] Es común denotar con $\pi(\theta)$ a la distribución $p(\theta)$ y, con $\pi(\theta \mid x)$ a $p(\theta \mid x)$.

Distribución Inicial Subjetiva

Consideramos primero la situación en la cual $\pi(\theta)$ es determinada, totalmente, de forma subjetiva. Si Θ es discreto, el problema se reduce a determinar la probabilidad subjetiva de cada elemento de Θ. Cuando Θ es continua, el problema de definir $\pi(\theta)$ es considerablemente más difícil.

1. *Histograma* [2, 29]

   Cuando Θ es un intervalo sobre la recta real, un enfoque empleado es el del *histograma*. Θ debe dividirse en intervalos determinando la probabilidad subjetiva de cada uno de ellos y entonces, representar el histograma de probabilidad. A partir del histograma puede esquematizarse una densidad $\pi(\theta)$ suave. Note que no existe una regla clara que establezca cuántos intervalos deben emplearse. Para algunos problemas es necesario emplear histogramas ordinarios mientras que para otros se necesitan versiones detalladas. La desventaja, de este enfoque, es la dificultad de trajabar con la densidad obtenida, además, aunque la probabiliad en los extremos de la distribución es frecuentemente pequeña, éstas pueden influenciar las inferencias subsecuentes.

2. *Verosimilitud Relativa* [29]

   Otro enfoque es el de la *verosimilitud relativa*, que también es empleado cuando Θ es un subconjunto de la recta real, siendo similar al enfoque anterior. Consiste simplemente en comparar la 'verosimilitud' intuitiva de varios puntos en Θ, trazando a partir de estas determinaciones una densidad inicial. Aún cuando $P(\theta = \theta_0) = 0$, $\forall \theta_0 \in \Theta$, esto puede realizarse, ya que:

   $$P(\theta = \theta_0 \mid \theta = \theta_0 \cup \theta = \theta_1) = \frac{P(\theta_0)}{P(\theta_0) + P(\theta_1)},$$

   donde $P(\theta)$ es la densidad inicial de θ. Así, un conjunto de valores proporcional a la densidad inicial de θ puede ser evaluado. La desventaja con este enfoque es la evaluación de la constante de normalización, ya que por construcción esta densidad no necesariamente es propia, en el sentido de que no integra uno; así mismo, cuando Θ es no acotado, ya que se tiene que determinar la forma de la densidad fuera de una región finita.

3. *Adaptación de una forma funcional dada* [2]

   Este enfoque es el más empleado. Consiste en asumir que $\pi(\theta)$ tiene una forma funcional dada, es decir, $\pi(\theta) \sim g(\theta \mid \lambda)$, con $\lambda \in \Lambda$ [19], es entonces necesario, sólo, elegir el hiperparámetro λ.

   Una forma de determinar subjetivamente el parámetro, es calcularlo a partir de los momentos iniciales estimados:

   $$\mu = E[\theta] = \int_{-\infty}^{\infty} \theta g(\theta \mid \lambda) d\theta = \phi(\alpha).$$

   $$\sigma^2 = V(\theta) = E[\theta - E[\theta]]^2 = \varphi(\alpha).$$

   Desafortunadamente, la estimación de los momentos iniciales, lleva al hecho de que los extremos de la densidad $g(\theta \mid \lambda)$ puden llevar a efectos drásticos sobre los momentos, si el parámetro es no acotado. Un mejor método para determinar parámetros iniciales subjetivamente es estimando varios percentiles de la distribución inicial y entonces, elegir los parámetros de la forma funcional dada para obtener una densidad que coincida con los percentiles estimados.

[19] λ es llamado *hiperparámetro*.

a) *Familia Conjugada* [2, 23, 29]

Se ha mencionado que el conocimiento sobre θ puede emplearse para especificar una distribución incial con una forma particular. Así, una familia paramétrica de densidades puede, en este caso, ser definida. Frecuentemente ésta familia hace que el análisis sea más fácil pero debe tenerse cuidado en elegir la densidad, esta debe realmente representar la información con la que se dispone. Formalmente una familia conjugada esta definiada por:

Definición 2.3.1. Sea $\mathcal{F} = \{p(x \mid \theta) : \theta \in \Theta\}$ una familia de distribuciones observacionales o muestrales. Una clase $\mathcal{P}$ de distribuciones es llamada una familia conjugada con respecto a $\mathcal{F}$:

$$\text{Si para todo } p(x \mid \theta) \in \mathcal{F} \text{ y } \pi(\theta) \in \mathcal{P} \text{ se tiene, } \pi(\theta \mid x) \in \mathcal{P}.$$

La tabla 2.2 muestra algunas distribuciones iniciales conjugadas.

$p(x \mid \theta)$	$\pi(\theta)$	$\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta)$
$\mathbf{Br}(x \mid \theta)$	$\mathbf{Be}(\theta \mid \alpha, \beta)$	$\mathbf{Be}(\theta \mid \gamma, \delta)$
$\mathbf{Pn}(x \mid \lambda)$	$\mathbf{Ga}(\theta \mid \alpha, \beta)$	$\mathbf{Ga}(\theta \mid \gamma, \delta)$
$\mathbf{Bi}(x \mid \theta, n)$	$\mathbf{Be}(\theta \mid \alpha, \beta)$	$\mathbf{Be}(\theta \mid \gamma, \delta)$
$\mathbf{N}(x \mid \mu, \sigma)$	$\mathbf{N}(\theta \mid \eta, \epsilon)$	$\mathbf{N}(\theta \mid \gamma, \delta)$
$\mathbf{N}(x \mid \mu, \sigma)$	$\mathbf{Ga}(\theta \mid \alpha, \beta)$	$\mathbf{Ga}(\theta \mid \gamma, \delta)$

Tabla 2.2: Iniciales Conjugadas

4. *Función de Distribución Acumulativa* [29]

Esta técnica es la menos empleada, aunque razonablemente más exacta es más fácil identificar una distribución a través de su densidad que a través de su función de distribución. En primer lugar se deben definir los percentiles. z_α es el $100\,\%\alpha$ percentil (α cuantil) de X si:

$$P(X \leqslant z_\alpha) = \alpha \text{ con } \alpha \in [0, 1].$$

Una vez definidos los percentiles, la colección de todos ellos decribe la función de distribución de X. En este enfoque, algunos percentiles son calculados subjetivamente como en el caso discreto.

De los cuatro enfoques presentados, el primero y segundo son los más utilizados para determinar subjetivamente una distribución inicial. El tercer enfoque es útil en dos situaciones. La primera, cuando una densidad de una forma funcional estándar coincide adecuadamente con la densidad obtenida por el primero o segundo enfoque y, la segunda, cuando se cuenta con información vaga sobre el parámetro y pueda incorporarse en la forma funcional seleccionada [2].

Generalmente se tiene que determinar una densidad inicial multivariante ($\Theta \subseteq \mathbb{R}^k$) más que una distribucíon univariante ($\Theta \subseteq \mathbb{R}^1$). El enfoque empleado es el uso de la forma funcional, dado que sólo algunos parámetros necesitan ser determinados subjetivamente, también cuando las coordenadas de $\theta = (\theta_1, \theta_2, ..., \theta_k)$ sean independientes, es decir:

$$\pi(\theta) = \prod_{i=1}^{k} \pi(\theta_i).$$

En caso contrario, es decir, cuando exista dependencia entre los parámetros la mejor forma para definir la distribución incial multivariante es a través de la determinación de las distribuciones

condicionales y marginales, es decir, se debe definir por técnicas univariantes las distribuciones condicionales iniciales $\pi(\theta_j \mid \theta_1, ..., \theta_{j-1})$ para $j = 1, ..., k$, para diferentes valores de $\theta_1, ..., \theta_{j-1}$, así:

$$\pi(\theta_1, ..., \theta_k) = \prod_{j=1}^{k} \pi(\theta_j \mid \theta_1, ..., \theta_{j-1}).$$

Hasta ahora, se ha supuesto que se cuenta con información inicial sobre el parámetro de interés, en caso contrario, se debe definir una densidad inicial que tenga una influencia mínima sobre la inferencia (*inicial no-informativa*).

Distribución Inicial No-Informativa

Lo atractivo de emplear las técnicas Bayesianas para realizar un análisis condicional, a pesar de no contar con información inicial sobre el parámetro de interés, es el hecho de aplicar, en tales situaciones, una distribución inicial no-informativa [10, 16]. Pero ¿cómo definir una inicial no-informtiva?.

El postulado de *Bayes/Laplace* establece que: *When nothing is known about θ in advance, let the prior $\pi(\theta)$ be a uniform distribution, that is, let all possibles outcomes of θ have the same probability* [20].

La situación más simple es entonces, cuando se considerea a Θ un conjunto finito. Si Θ consiste de n elementos entonces, la inicial no-informativa más obvia es la que asigna a cada elemento de Θ la probabilidad de $\frac{1}{n}$. Esto puede generalizarse por supuesto; si Θ es un conjunto infinito, se puede asignar a cada $\theta \in \Theta$ la misma densidad, obteniendo así, la inicial no-informativa uniforme $\pi(\theta) = c$. El principal problema de emplear la distribución uniforme como inicial no-informativa, es que la distribución uniforme es no invariante bajo reparametrizaciones, es decir, si no se cuenta con información sobre θ entonces, tampoco sobre $\phi = g(\theta)$[20] así:

$$\text{Si } \pi(\theta) = c \text{ entonces, } \phi = g(\theta) \Rightarrow \pi(\phi) = |\tfrac{d}{d\phi} g^{-1}(\phi)| \cdot \pi(g^{-1}(\phi)).$$

También cuando el espacio paramétrico Θ es infinito, la incial uniforme presenta el problema de ser impropia, aunque no siempre es serio este problema, iniciales impropias, frecuentemente, llevan a posteriores propias.

1. *Inicial invariante* [20]

   Este enfoque busca una estructura invariante en el problema, dejando que la distribución inicial tenga la misma estructura invariante. Matemáticamente, el modelo y la distribución inicial deben ser invariantes bajo la acción del mismo grupo empleando la medida de *Haar derecha*[21] como la inicial no informativa; ésta medida tiene la característica de ser invariante por multiplicaciones a la derecha con el grupo pero una dificultad con este enfoque, es que todos los problemas no tienen una estructura invariante, además, la medida Haar derecha no siempre existe.

[20] $g(\theta)$ función uno-a-uno.
[21]

Definición 2.3.2. Sea G un grupo topológico localmente compacto. Sea X la σ-álgebra generada por los subconjuntos compactos de G (álgebra de Borel). Sea $a \in G$ y $S \subseteq G$ y defina la traslación derecha de S como: $Sa = \{sa : s \in S\}$. Una medida μ sobre los subconjuntos de Borel de G es invariante por traslación derecha si y sólo si $\mu(Sa) = \mu(S)$ para todo $S \subset G$ y toda $a \in G$. Sea S^{-1} el conjunto de elementos inversos S, si $\mu(S^{-1})$ es invariante por traslación derecha entonces, μ es llamada la medida derecha de Haar.

2. *Inicial de Jeffreys* [2, 15, 21, 27, 33, 43]

   Este método es el más empleado y consiste en elegir como la inicial no-informativa a:

$$\pi(\theta) = [I(\theta)]^{\frac{1}{2}},$$

   donde $I(\theta)$ es la información esperada de Fisher, la cual, bajo suposiciones comunes está dada por:

$$I(\theta) = -E_\theta[\frac{\partial^2 log f(X \mid \theta)}{\partial \theta^2}].$$

   Si $\theta = (\theta_1, ..., \theta_k)^t$, la regla de Jeffreys sugiere como inicial no-informativa el empleo de $\pi(\theta) = [detI(\theta)]^{\frac{1}{2}}$, donde $I(\theta)$ es la matriz $(k \times k)$ de información de Fisher, dada por:

$$I_{(i,j)} = -E_\theta[\frac{\partial^2}{\partial \theta_i \partial \theta_j} log f(X \mid \theta)].$$

3. *Inicial de Referencia* [3, 4, 13, 37]

   La inicial de referencia, es otra clase de inicial no-informativa y cuya idea clave es la *información esperada*[22].

   Definición 2.3.3. Se define la información esperada de una observación a partir del modelo $\mathcal{M} = \{p(x \mid \theta) : x \in \mathcal{X}, \theta \in \Theta\}$, cuando $q(\theta)$ es la distribución inicial para θ, como:

$$I\{q \mid \mathcal{M}\} = \int \int_{\mathcal{X} \times \Theta} p(x \mid \theta) q(\theta) log \frac{p(\theta \mid x)}{q(\theta)} dx d\theta = \int_{\mathcal{X}} kq \mid p(\cdot \mid x) p(x) dx,$$

   donde $p(\theta \mid x) = \frac{p(x|\theta)q(\theta)}{p(x)}$ y $p(x) = \int_\theta p(x \mid \theta) q(\theta) d\theta$.

   Considerese ahora la información $I\{q \mid \mathcal{M}^k\}$, la información esperada de k repeticiones independientes de $\mathcal{M}$. Cuando $k \to \infty$, la secuencia de realizaciones $\{x_1, ..., x_k\}$ finalmente proporciona toda la información faltante sobre el valor de θ. Por lo tanto, cuando $k \to \infty$, $I\{q \mid \mathcal{M}^k\}$ proporciona una medida de la información faltante sobre θ asociada a la incial $q(\theta)$. Intuitivamente, una inicial de referencia será una inicial permisible[23] que maximiza la falta de información sobre θ dentro de la clase $\mathcal{P}$ de inciales compatibles con cualquier conocimiento asumido sobre el valor de θ.

   Algunas dificultades que presenta la información esperada es, dentro de los espacios paramétricos continuos, ya que la información faltante $I\{q \mid \mathcal{M}^k\}$ generalmente diverge cuando $k \to \infty$, debido a que una cantidad infinita de información sería necesaria para conocer el verdadero valor de θ, así mismo, la información esperada no está definida sobre un conjunto no acotado. De lo anterior, se hace necesario entonces, la siguiente definición:

[22] Información que depende de una distribución inicial.
[23]

Definición 2.3.4. Una función continua estrictamente positiva, $\pi(\theta)$, es una incial permisible para el modelo $\mathcal{M}$ si:

a) $\pi(\theta \mid x)$ es propia para toda $x \in \mathcal{X}$ y

b) Para una aproximación secuencial compacta, la correspondiente secuencia posterior converge logaritmicamente a $\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta)$.

Definición 2.3.5. Sea $\mathcal{M} = \{p(x \mid \theta) : x \in \mathcal{X}, \theta \in \Theta \in \Re\}$, un modelo con un parámetro continuo y sea $\mathcal{P}$ una clase de funciones inciales sobre θ para las cuales $\int_{\Theta} p(x \mid \theta)p(\theta)d\theta < \infty$. La función $\pi(\theta)$ se dirá que tiene la *propiedad de maximización de la información faltante* para el modelo $\mathcal{M}$ dado $\mathcal{P}$ si, para cualquier conjunto compacto[24] $\Theta_0 \in \Theta$ y cualquier $p \in \mathcal{P}$:

$$\lim_{k \to \infty} \{I\{\pi_0 \mid \mathcal{M}^k\} - I\{p_0 \mid \mathcal{M}^k\}\} \geq 0,$$

donde π_0 y p_0 son las restricciones renormalizadas de $\pi(\theta)$, respectivamente, y $p(\theta)$ en Θ_0.

Así, una incial de referencia queda definiada por:

Definición 2.3.6. Una función $\pi(\theta) = \pi(\theta \mid \mathcal{M}, \mathcal{P})$ es una incial de referencia para el modelo $\mathcal{M}$ dado $\mathcal{P}$ si es permisible y tiene la propiedad de la maximización de la información faltante.

Hasta aquí, se han presentado diferentes maneras para definir una distribución inicial, desde métodos puramente subjetivos, hasta enfoques indirectos, vía formas funcionales. Por supuesto, que una distribución inicial partícular debe ser elegida de acuerdo a la información disponible y al problema de inferencia a tratar, aunque las distribuciones iniciales no-informativas son las más empleadas no necesariamente ayudan a resolver cualquier problema inferencial.

2.4. Necesidad de los métodos de simulación

La mayor parte del esfuerzo en los procedimientos Bayesianos, se concentra en el cálculo de ciertas características (media, mediana, moda) de la distribución posterior del parámetro de interés [34]. Desafortunadamente, en la práctica $\pi(\theta \mid x)$ puede llegar a ser tan compleja que, generalmente, las integrales que implica pueden no resolverse analíticamente, haciendo necesario el empleo de métodos numéricos eficientes que permitan calcular, o aproximar, integrales en varias dimensiones.

Algunos métodos para calcular la distribución posterior son:

- Métodos Asintóticos (Aproximación normal y método de Laplace), empleados para distribuciones posteriores unimodales.
- Métodos Monte Carlo no iterativos (muestreo directo y muestreo indirecto).
- Métodos MCMC (Metrópolis-Hastings y Muestreador de Gibbs).

Las técnicas de Monte Carlo han mostrado ser las más flexibles. La implementación de las técnicas de Monte Carlo vía Cadenas de Morkov (MCMC) permiten generar, de manera iterativa, observaciones de distribuciones multivariadas que difícilmente podrían simularse utiliazando métodos directos. Los dos algoritmos más empleados para construir cadenas de Markov con las propiedades deseadas son el de Metrópolis-Hastings y el muestreo de Gibbs, este último ha dado un impulso extraordinario a las aplicaciones de las técnicas Bayesianas.

El método Muestreador de Gibbs parte de los valores iniciales asignados a los parámetros y va generando valores de los mismos, en el primer paso obtiene el del primer parámetro, teniendo en cuenta los valores iniciales de los restantes, en el segundo paso obtiene la estimación del segundo parámetro teniendo en cuenta la estimación actual del primer parámetro y los valores iniciales de los restantes, y así sucesivamente hasta formar una cadena de Markov.

La idea básica es sencilla: *construir una cadena de Markov que sea fácil de simular y cuya distribución de equilibrio corresponda a la distribución final de interés* [17].

[24] La restricción a un conjunto compacto, generalmente asegura la existencia de la falta de información para un k dado.

El propósito de esta sección es presentar algunas técnicas avanzadas de simulación. Dentro de la simulación estocástica, se encuentra la simulación basada en las técnicas de Monte Carlo, las cuales son útiles para estudiar procesos estocásticos, sistemas determinísticos que involucran modelos analíticos muy complicados y problemas determinísticos de muchas dimensiones cuya discretización no es factible, como ocurre con problemas de integración. Los avances computacionales, así como el mejoramiento de los lenguajes de programación han permitido el desarrollo de las técnicas de simulación y por ende, la simulación Monte Carlo basada en cadenas de Markov.

Método de Monte Carlo

El método Monte Carlo se utiliza para obtener una aproximación numérica de integrales cuyo valor no es inmediato. La idea básica es calcular una integral en términos del valor esperado de alguna función con respecto a alguna distribución de probabilidad. Para ilustrar el método, suponga que desea calcular la integral de una función f(x), continua en el intervalo [a, b]; es decir:

$$I = \int_a^b f(x)dx.$$

Para encontrar un valor aproximado a I se requiere de una función de densidad g(x), definida en [a, b]. Entonces:

$$\int_a^b f(x)dx = \int_a^b f(x)\frac{g(x)}{g(x)}dx.$$

A través de esta expresión se reconoce el valor esperado de la función $h(x) = \frac{f(x)}{g(x)}$, por medio del cual se obtiene el valor de I, es decir:

$$I = \int_a^b f(x)\frac{g(x)}{g(x)}dx = \int_a^b h(x)g(x)dx = E[h(x)].$$

Para obtener este valor esperado es posible simular variables aleatorias independientes e idénticamente distribuidas $X_1, X_2, ..., X_n$, con función de densidad $g(x)$. Por la ley fuerte de los grandes números:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = I.$$

Por lo tanto:

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

Es un estimador consistente e insesgado de I.

Métodos Computacionales vía Cadenas de Markov

Una cadena de Markov es un proceso estocástico $\{\theta^t : t \in T\}$ tal que, dado el estado presente, el pasado y el futuro del proceso son independientes [35]. En otras palabras:

$$P(\theta^{(n+1)} \in E|\theta^{(n)} \in E_n, \theta^{(n-1)} \in E_{n-1}, ..., \theta^{(0)} \in E_0) = P(\theta^{(n+1)} \in E|\theta^{(n)} \in E_n)$$

para todos los conjuntos $E_0, ..., E_{n-1} \subset \sigma(S)$, donde S denota el espacio de estados de la cadena, con S-finito, y $\sigma(S)$ es una σ-álgebra de subconjuntos de S.

Los métodos Monte Carlo Vía cadenas de Markov (MCMC) permiten generar, de manera iterativa, observaciones de distribuciones multivariadas que difícilmente podrían simularse utilizando métodos directos. La idea básica es construir una cadena de Markov que sea fácil de simular y cuya distribución de equilibrio corresponda a la distribución final de interés; de tal forma que los métodos MCMC requieren que una cadena sea:

1. Homogénea, es decir, que $P(\theta^{(n+1)} \in E|\theta^{(n)} \in E_n)$ no dependa de n.

2. Irreducible, desde cualquier estado se puede acceder a otro, todos los estados se comunican entre sí.

3. Aperiódica.

4. Reversible.

Es importante mencionar que la propiedad de reversibilidad asegura la existencia de una cadena ergódica. Cuando el espacio de estados E es numerable, esta propiedad se sigue del hecho de que la cadena sea homogénea en el tiempo, irreducible y aperiódica. Sin embargo, cuando el espacio de estados E es no numerable, que es el caso más común en aplicaciones del algoritmo de Metropolis-Hastings y muestreo de Gibbs, existen otras consideraciones, consultar [35].

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

1. $\theta^{(t)} \to^{\mathcal{D}} \theta$, donde $\theta \sim p(\theta|x)$.

2. $\frac{1}{t}\sum_{i=1}^{t} g(\theta^{(i)}) \to E[g(\theta|x)]$ c. s.

Cabe señalar que a pesar de que los valores iniciales influyen en el comportamiento de la cadena, conforme el número de iteraciones aumenta, los valores iniciales pierden importancia y el proceso converge a una distribución de equilibrio. Por eso, en la aplicación es común que las iteraciones iniciales sean descartadas. El problema por lo tanto consiste en construir algoritmos que generen cadenas de Markov cuya distribución de equilibrio coincida con la distribución de interés, algunos algoritmos que abordan esta problemática son el algoritmo de Metropolis Hasting y el algoritmo de Gibbs.

Metropolis-Hastings

Este algoritmo permite construir una cadena de Markov al definir probabilidades de transición de la siguiente manera.

Sea $Q(\theta^*|\theta)$ una distribución de transición (arbitraria) y se define:

$$\alpha(\theta^*, \theta) = min\{\frac{p(\theta^*|x)Q(\theta|\theta^*)}{p(\theta|x)Q(\theta^*|\theta)}, 1\}.$$

La idea es generar un valor de una distribución auxiliar y aceptarla con una probabilidad dada. Este mecanismo de corrección garantiza la convergencia de la cadena a su distribución de equilibrio. Suponga que la cadena está en el estado θ y se genera un valor θ^* de una distribución opuesta $Q(\cdot|\theta)$. Un nuevo valor θ^* es aceptado con probabilidad α. La cadena puede permanecer en un mismo estado a lo largo de muchas iteraciones. En la práctica, se acostumbra monitorear esto para calcular el porcentaje medio de iteraciones para los cuales los nuevos valores se aceptan.

El algoritmo de Metropolis-Hastings se especifica de la siguiente manera:

Dado un valor inicial $\theta^{(0)}$, la t-ésima iteración consiste en:

1. Generar una observación θ^* de $Q(\theta^*|\theta^{(t)})$.

2. Generar una variable $u \sim U(0, 1)$.

3. Si $u \leq \alpha(\theta^*, \theta^{(t)})$, hacer $\theta^{(t+1)} = \theta^*$; en caso contrario, $\theta^{(t+1)} = \theta^{(t)}$.

Este procedimiento genera una cadena de Markov con distribución de transición:

$$p(\theta^{(t+1)}|\theta^{(t)}) = \alpha(\theta^{(t+1)}, \theta^{(t)})Q(\theta^{(t+1)}|\theta^{(t)}).$$

La probabilidad de aceptación $\alpha(\theta^*, \theta)$ sólo depende de $p(\theta|x)$ a través de un cociente, por lo que la constante de normalización no es necesaria [25]. En la práctica, frecuentemente se utiliza alguno de los dos siguientes casos:

- Caminata Aleatoria. Sea $Q(\theta^*|\theta) = Q_1(\theta^* - \theta)$, donde $Q_1(\cdot)$ es una densidad de probabilidad simétrica centrada en el origen. Entonces:

$$\alpha(\theta^*, \theta) = min\{\frac{p(\theta^*|x)}{p(\theta|x)}, 1\}.$$

- Independencia. Sea $Q(\theta^*|\theta) = Q_0(\theta^*)$, donde $Q_0(\cdot)$ es una densidad de probabilidad sobre Θ, por tanto:

$$\alpha(\theta^*, \theta) = min\{\frac{\omega(\theta^*)}{\omega(\theta)}, 1\}$$

  con $\omega(\theta) = p(\theta|x)Q_0(\theta)$.

Es común emplear, después de una reparametrización apropiada, distribuciones de tansición normales o t de Student ligeramente sobres dispersas:

$$Q(\theta^*|\theta) = N_d(\theta^*|\theta, \kappa V(\hat{\theta})) \text{ (Caminata Aleatoria)},$$

$$Q_0(\theta^*) = N_d(\theta^*|\hat{\theta}, \kappa V(\hat{\theta})) \text{ (Indenpendencia)},$$

donde $\hat{\theta}$ y $V(\hat{\theta})$ denotan a la media y a la matriz de varianzas-covarianzas de la aproximación asintótica normal para $p(\theta|x)$, respectivamente y, $\kappa \geq 1$ es un factor de sobredispersión.

Algoritmo de Gibbs

El algoritmo de Gibbs permite simular una cadena de Markov $\theta^{(1)}, \theta^{(2)}, ...$ con distribución de equilibrio $p(\theta|x)$. Cada valor nuevo de la cadena se obtiene al generar muestras de distribuciones cuya dimesión es menor que d y que en la mayoría de los casos tiene una forma más sencilla que la de $p(\theta|x)$.

Sea $\theta = (\theta_1, ..., \theta_n)$ una partición del vector θ, donde $\theta_i \in \Re^{d_i}$ y $\sum_{i=1}^{n} d_i = d$. En este caso θ_i es un vector, pero en general cada componente θ_i puede ser un escalar, un vector o una matriz. Las siguientes densidades para cada θ_i son conocidas como *densidades condicionales completas*:

$$\begin{array}{c} p(\theta_1|\theta_2, ..., \theta_n, x) \\ . \\ . \\ . \\ p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n, x) \text{ para toda } i = 2, ..., n-1 \\ . \\ . \\ . \\ p(\theta_n|\theta_1, ..., \theta_{n-1}, x). \end{array}$$

[25] La versión original del algoritmo de Metropolis toma a Q tal que $Q(\theta^*|\theta) = Q(\theta|\theta^*)$, en cuyo caso $\alpha(\theta^*, \theta) = min\{\frac{p(\theta^*|x)}{p(\theta|x)}, 1\}$.

Estas densidades pueden identificarse fácilmente al inspeccionar la forma de la distribución final $p(\theta|x)$. De hecho para cada $i = 1, 2, ..., n$:

$$p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n, x) \propto p(\theta|x)$$

donde $p(\theta|x)$ es vista sólo como función de θ_i.

Dado un valor inicial $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_n^{(0)})$, el algoritmo de Gibbs simula una cadena de Markov en la que $\theta^{(t+1)}$ se obtiene a partir de $\theta^{(t)}$ del siguiente modo:

generar una observación $\theta_1^{(t+1)}$ de $p(\theta_1|\theta_2^{(t)}, ..., \theta_n^{(t)}, x)$;

generar una observación $\theta_2^{(t+1)}$ de $p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_n^{(t)}, x)$;

generar una observación $\theta_3^{(t+1)}$ de $p(\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, ..., \theta_n^{(t)}, x)$;
generar una observación $\theta_4^{(t+1)}$ de $p(\theta_4|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_5^{(t)}, ..., \theta_n^{(t)}, x)$;

.
.
.

generar una observación $\theta_n^{(t+1)}$ de $p(\theta_n|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_3^{(t+1)}, ..., \theta_{n-1}^{(t+1)}, x)$.

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

$$p(\theta^{(t+1)}|\theta^{(t)}) = \prod_{i=1}^{n} p(\theta_i^{(t+1)}|\theta_1^{(t+1)}, ..., \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, ..., \theta_n^{(t)}, x).$$

2.4.1. Softwares estadísticos Bayesianos

Los procedimientos Bayesianos son fáciles siempre y cuando se apeguen a distribuciones bien conocidas, pero son muy pocos y resultan ser para los modelos más simples. Sólo para algunos modelos se puede aproximar o incluso calcular analíticamente la distribución final, por lo que, la implementación de las técnicas Bayesianas requieren de un esfuerzo computacional alto. El desarrollo y mejoramiento de las técnicas Monte Carlo basadas en cadenas de Markov (MCMC) y la construcción de softwares de cómputo de propósito general, han hecho posible construir (o aproximar) la distribución final de modelos Bayesianos complejos.

La presente sección expone, brevemente, el desarrollo de la estadística Bayesiana computacional y el procedimiento básico a seguir para analizar un modelo Bayesiano con determinada herramienta de software.

Reseña Histórica

Aproximadamente hace 20 años, la estadística Bayesiana computacional moderna surge al implementar los métodos MCMC con el desarrollo, independiente, de los paquetes informáticos **BUGS** (Bayesian inference Using Gibbs Sampling)[26] y **JAGS** (Just Another Gibbs Sampler) [1, 39].

Más recientemente, se ha adaptado **OpenBUGS**[27] al programa **R**[28], la modelación y la flexibilidad de inferencia de OpenBUGS en combinación con el poder de manipulación de datos y gráfico

[26]BUGS ha pasado a denominarse WinBUGS y, más recientemente, OpenBUGS.
[27]**rbugs** (**BRugs**) es la adaptación de OpenBugs a R y **R2WinBugs** de WinBUGS.
[28]Software estadístico libre: http://cran.opensourceresources.org.

de R, permite explorar las muestras MCMC obtenidas con OpenBUGS. Aunque esta adaptación es una aplicación orientada a instrucciones, existen paquetes (aplicaciones) orientados hacia modelos y paquetes bajo un entorno totalmente Bayesiano. Algunas de estos paquetes son: **MCMCpack**, **Zelig** (aplicaciones orientadas hacia modelos), **rv** (simulación basada en variables aleatorias), **Umacs** (Universal Morkov Chain Sampler), **CODA** (Convergence Diagnosis and Output Analysis), **BOA** (Bayesian Output Analyisis program), **BMA** (Bayesian Model Averangin), entre otros paquetes[29].

Paquetes Informáticos

El análisis estadístico Bayesiano, dependerá esencialmente del paquete informático a emplear; aunque varios de los paquetes mencionados comparten ciertas funciones, cada uno de ellos está enfocado a realizar un análisis específico, por ejemplo, CODA y BOA están diseñados para completar y validar los resultados generados por los procedimientos MCMC, mientras que BMA esta diseñado para validar aspectos específicos bajo el enfoque Bayesiano.

A continuación, y de manera *esquemática*, se presenta el procedimento básico para analizar un modelo Bayesiano bajo los siguientes softwares:

- **WinBUGS**.
- **R** (empleando los paquetes):
  - MCMCpack.
  - Zelig.

WinBUGS

WinBUGS[30] proporciona herramientas simples para llevar a cabo la simulación Monte Carlo, llevando a cabo el muestreo de Gibbs para una gran variedad de modelos cuya descripción se hace por medio del lenguaje BUGS o a través de Doodles (representación gráfica de modelos). Los modelos en WinBUGS se pueden traducir, casi literalmente, de la descripción escrita usual al lenguaje BUGS, aunque, tiene ciertas limitaciones para detectar la convergencia, resumir las muestras y realizar diagnósticos sobre el ajuste de los modelos. Tiene, sin embargo, una característica, funciona como un sistema experto, es decir, intenta usar el método de muestreo más apropiado para cada nodo estocástico (Ver Tabla 2.3)[17].

Distribución Objetivo	Método de Muestreo
Discreta	Inversión de una función de distribución acumulativa
Forma Cerrada	Muestreo directo usando algoritmos estándar
Log-cóncava	Muestreo de rechazo adaptativo (Gilks 1992)
Rango restringido	Muestreo slice (Neal 1997)
Rango sin restricción	Metropolis-Hastings (Metropolis et al. 1953, Hastings 1970)

Tabla 2.3: Método de muestreo apropiado.

[29]Descargables desde el CRAN (The Comprehensive R Archive Network) en R.
[30]Software libre: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

El análisis de un modelo Bayesiano con WinBUGS, de manera global, es a través de los siguientes pasos:

- Especificación del modelo a analizar. (Ver Figura 2.1)
- Verificación de la sintáxis del modelo.
- Ingreso de datos.[31]

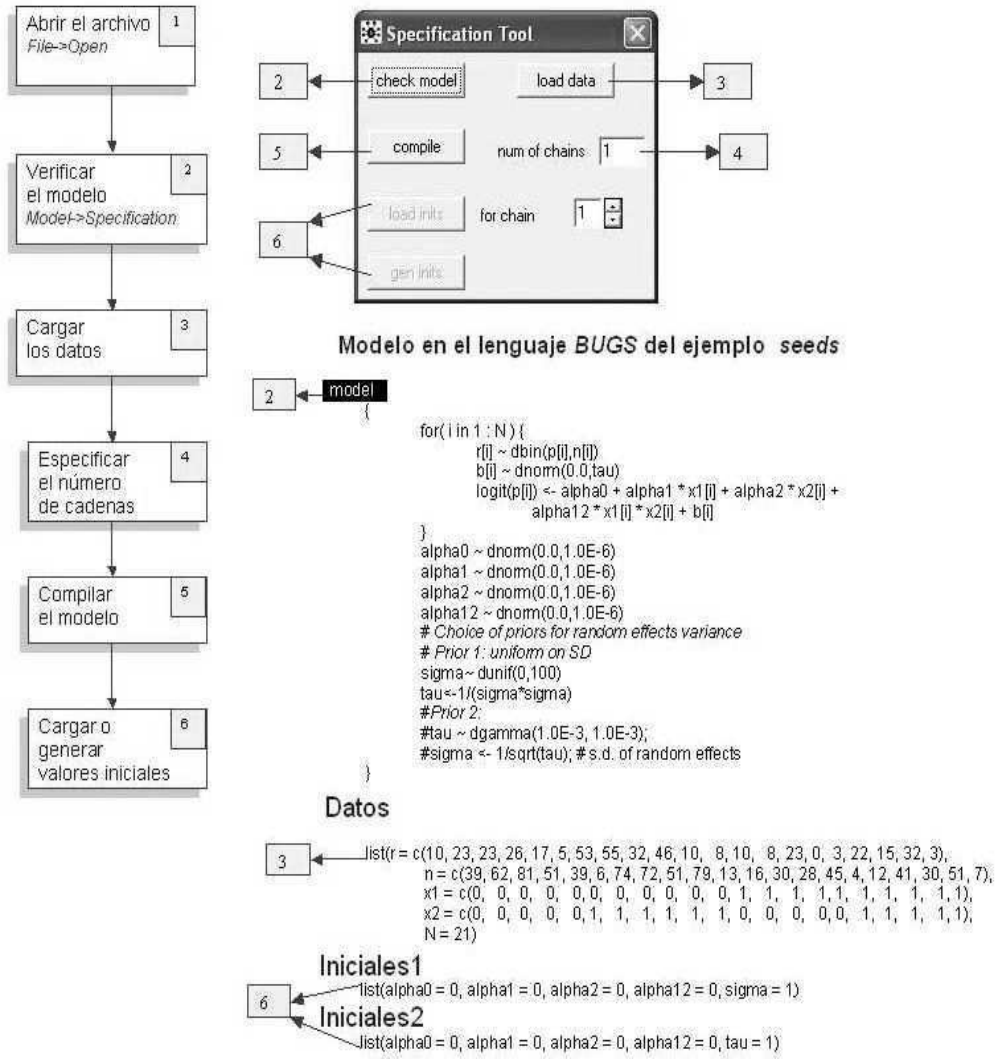


Figura 2.1: Especificación del Modelo

[31]Los datos pueden estar cargados como un objeto en **R**.

- Especificación del número de cadenas- muestras a simular. (Ver Figura 2.2)
- Compilación del modelo.
- Especificación de los valores iniciales para cada nodo.

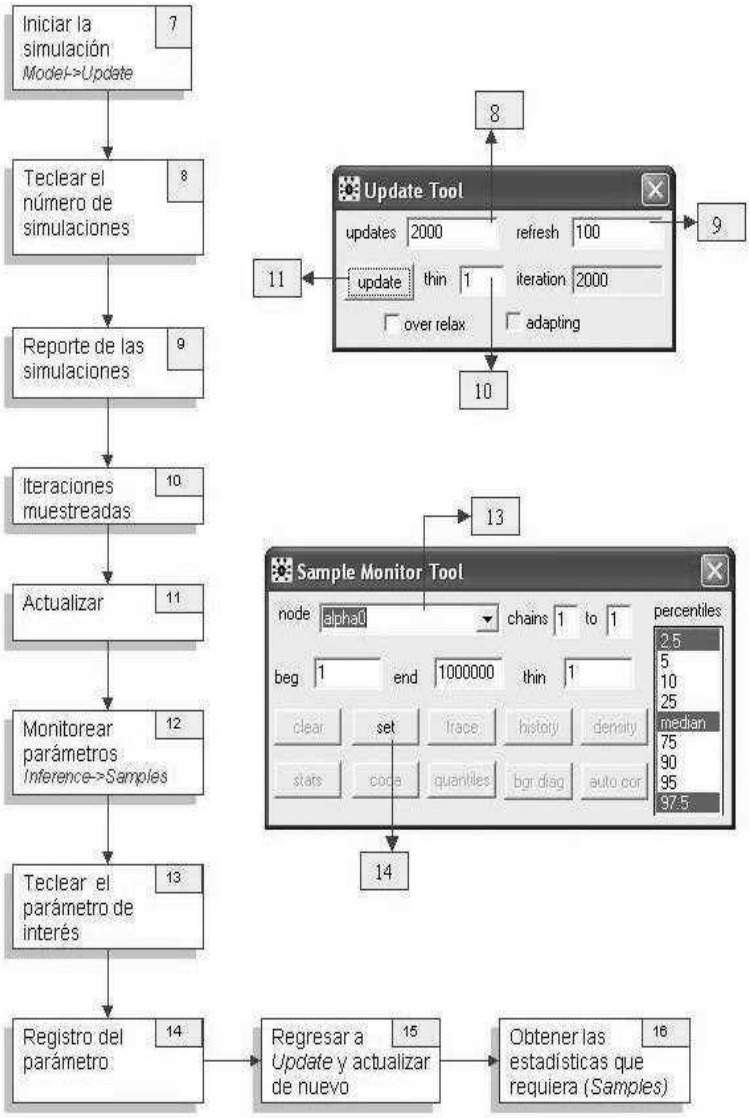


Figura 2.2: Monitoreo y obtención de las estadísticas de los parámetros

MCMCpack

MCMCpack contiene funciones para realizar inferencia Bayesiana para una serie de modelos estadísticos. Específicamente, MCMCpack analiza el modelo de regresión lineal normal con la función *MCMCregress*, esta función genera una muestra de la distribución posterior a través del muetreo Gibbs considerando por defecto, una distribución inicial normal multivariante para el vector de coeficientes y una distribución incial gamma inversa para la varianza no condicional [28].

Sintaxis MCMCpack:

MCMCregress(formula, data = NULL, burnin = 1000, mcmc = 10000, b0 = 0, B0 = 0, c0 = 0.001, ...)[32]

El análisis básico de un modelo Bayesiano con MCMCpack, es a través de los siguientes pasos:

- Ingreso de datos. (Ver Figura 2.3)
- Activación de la libreria MCMCpack.

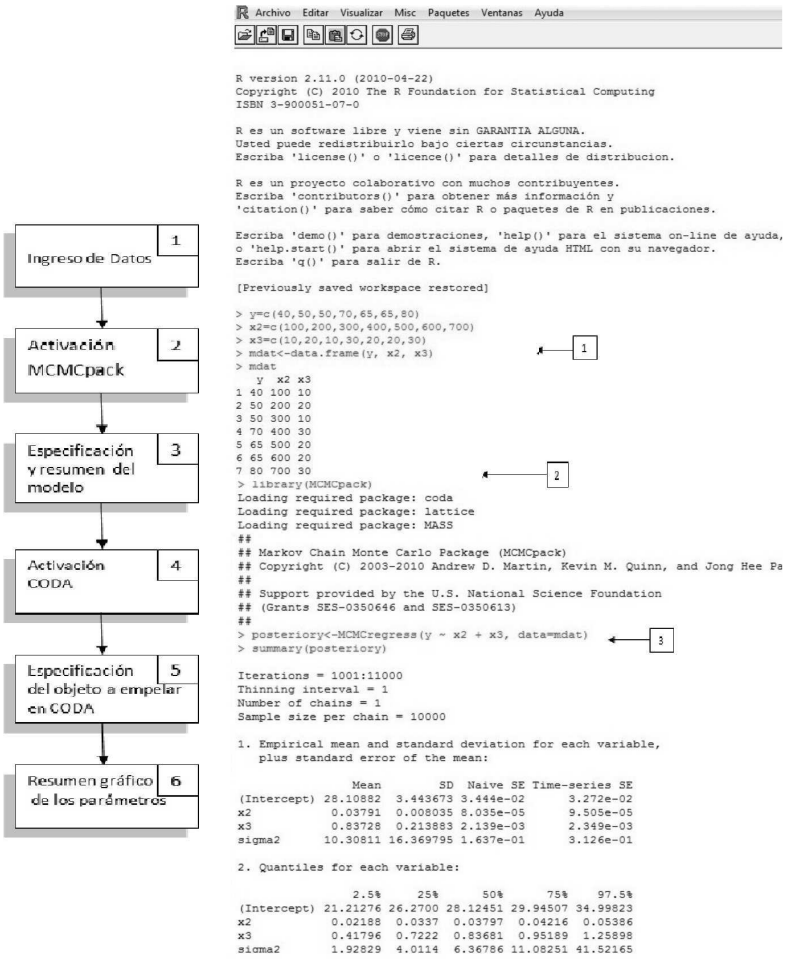


Figura 2.3: Activación y especificación del modelo en MCMCpack

[32]Una especificación más detallada sobre los argumentos puede consonsultarse en [28].

- Especificación del modelo
- Activación del paquete CODA. (Ver Figura 2.4)
- Resumen del modelo.

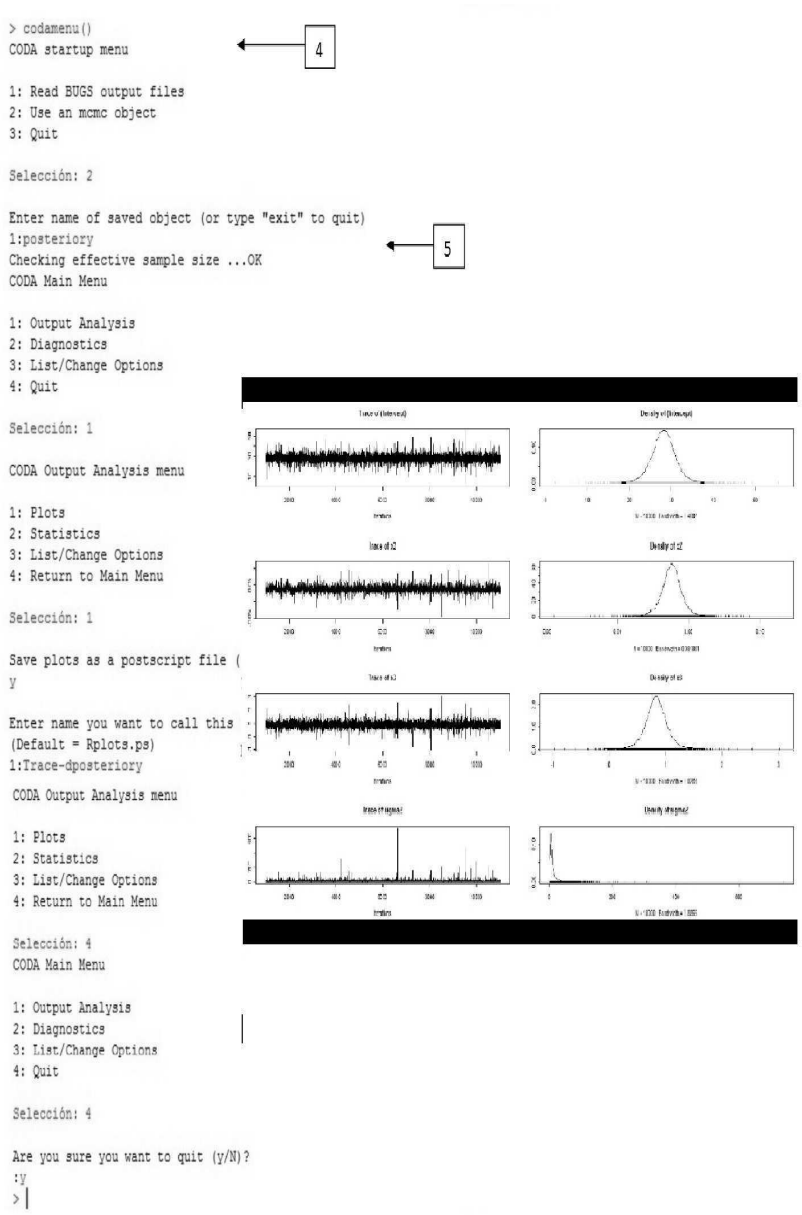


Figura 2.4: Resúmen estadístico con CODA

Observese que al solicitar el resumen del modelo (Fig.2.3) no sólo un valor único de los parámetros estimados es mostrado, sino, una muestra de la distribución final de los mismos incluyendo sus momentos: media, desviación estándar y cuantiles (con diversos porcentajes); todas estas estimaciones son obtenidas sobre una cadena de muestras, que por defecto tiene una longitud igual a 10000 valores. La simulación generada, retorna objetos *mcmc* CODA, los cuales pueden ser visualizados y analizados con el paquete CODA[33] (Fig. 2.4), obteniendo otra serie de resultados, como la representación gráfica de las distribuciones finales de los parámetros, las trazas de los valores simulados, así como diversos criterios de convergencia.

Zelig

Zelig es un programa simple y fácil de usar, ayudando a calcular, interpretar y presentar los resultados de una amplia gama de modelos estadísticos[34]. Sintaxis Zelig:

zelig(formula, model, data, by, save.data, cite, ...)[35]

El paquete informático Zelig, permite realizar estimaciones máximo verosímiles desde el enfoque clásico. La especificación *model = 'normal'*, analiza el modelo de regresión lineal normal, variante más cercana al modelo estándar de regresión por mínimos cuadrados[36], calculando, además de las estimaciones de los coeficientes del modelo de regresión considerado, valores esperados, así como sus primeras diferencias junto con sus medidas de incertidumbre asociadas a ellos (errores estándar e intervalos de confianza).

La estimación del modelo de regresión lineal normal Bayesiano, en particular, es determinada por la especificación *model = 'normal.bayes'*, tal especificación emplea una distribución inicial semi-conjugada para los parámetros (misma que la empleada por MCMCpack), además, entre otras opciones bajo la especificación del modelo normal Bayesiano, Zelig permite realizar pruebas de diagnóstico sobre la convergencia de la cadena MCMC a su distribución estacionaria, entre otras [22].

El análisis básico de un modelo con el paquete Zelig, es a través de los siguientes pasos:

- Ingreso de datos. (Ver Figura 2.5)
- Activación de la libreria Zelig.
- Definición del modelo y de las variables exploratorias.
- Diagnóstico de convergencia. (sólo en la generación de cadenas MCMC)
- Resumen analítico del modelo.
- Solicitud gráfica de los parámetros de interés.

[33] Paquete incorporado en la memoria al activar MCMCpack.
[34] *help.zelig("models")* ofrece una lista completa de modelos soportados por Zelig.
[35] La especificación de los argumentos puede consonsultarse en [22].
[36] Ambos modelos difieren, sólamente, por su estimación del parámetro alteatorio σ.

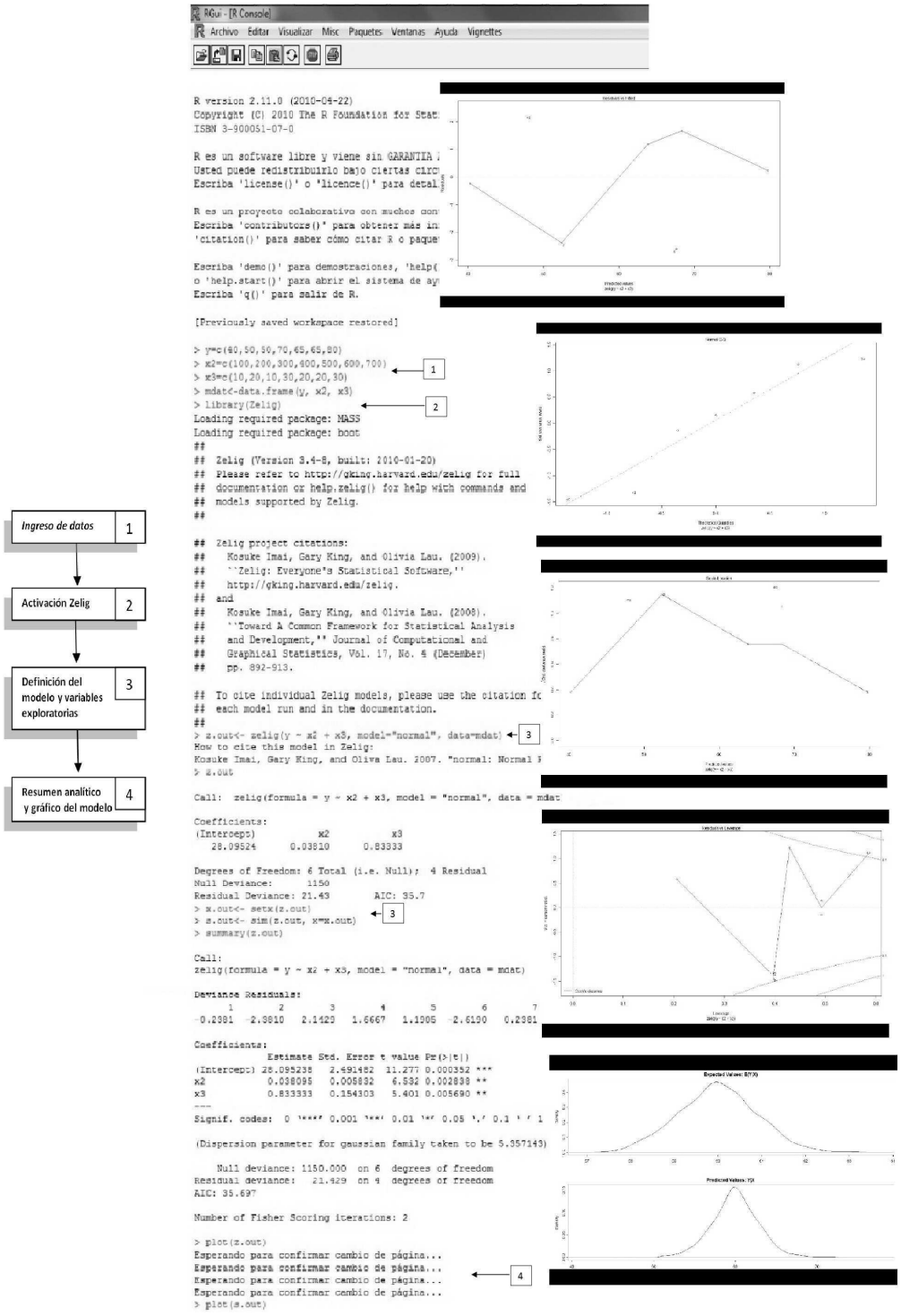


Figura 2.5: ZELIG

Capítulo 3

Ajuste Bayesiano de Curvas

Las razones por las que se puede estar interesado en la estimación de una curva son, primero, que provee una forma para explorar y presentar la relación entre la(s) variable(s) independiente(s) y la variable respuesta, y segundo, permite la predicción de observaciones futuras. Además, el método para estimar una curva también puede ser utilizado para dar estimaciones sobre propiedades de interés de la curva.

Las técnicas de regresión [30], se encuentran entre los métodos más extensamente empleados para dicho fin. Un interés es la estimación de una función de regresión $g(x) = E(y \mid x)$ basada sobre datos muestrales $z_i = (y_i, x_i)$ $(i = 1, 2, ..., n)$; donde y y x pueden ser escalares o vectores. Generalmente g es elegida con respecto a algún criterio que mide la cercanía de $g(x)$ a y. Por ejemplo, un tal criterio es el *error cuadrático medio*, $E[\{y - g(x)\}^2 \mid x]$, el cual es mínimo cuando $g(x) = E(y \mid x)$. Así, si se puede estimar la distribución conjunta de (y, x), o más directamente la distribución condicional de $(y \mid x)$, el problema es solucionado en principio por el cálculo de $g(x)$. Un enfoque general para el ajuste de curvas es permitir que g sea una función *polinomial a trozos* compuesta de polinomios de orden no cero, excepto en los puntos adyacentes a un nodo. Otra herramienta, frecuentemente empleada para la estimación de una curva, es el suavizamiento por splines; el suavizamiento por splines emplea algunos nodos localizados en los datos, u otra alternativa es el empleo del enfoque, frecuentemente llamado, *ajuste de curvas a través de splines con nodos libres*, debido a que el número de nodos y sus localizaciones son determinadas libremente, a partir de los datos. Existen algunos otros métodos para modelar la superficie de regresión, $E(Y \mid X)$, de una variable respusta Y, dada una matriz de covariables o predictores $X \in \Re^p$. Cada método tiene sus ventajas y desventajas. Sin embargo, existe una compensación común generalmente entre la flexibilidad del modelo y su interpretabilidad.

Los modelos lineales son sumamente interpretables pero resultan muy rígidos cuando se tiene una gran cantidad de datos y/o cuando una dependencia no-lineal existe. Los modelos aditivos generalizados hacen más flexibles las suposiciones lineales pero restringen los efectos no-lineales introducidos a través de cada covariante independientemente. Una alternativa, que es tanto interpretable como flexible, es proporcionada por el uso de modelos lineales locales y modelos lineales a trozos.

Los enfoques no-paramétricos también son empleados para ésta clase de problemas. La principal suposición entre estos métodos es que, si $x_1, x_2, ..., x_n$ son puntos designados de un experimento en el cual las observaciones $y_1, ..., y_n$ son obtenidas, existe una función de regresión desconida $g(x)$ la cual satisface $y = g(x) + \varepsilon(x)$, donde $g(\cdot)$ y $\varepsilon(\cdot)$ son las componentes sistemática y residual, respectivamente. La modelación debe establecer un equilibrio razonable entre suavizamiento y flexibilidad dentro de la estructura local y los cambios en la aparente relación de regresión. Una idea simple, que es la base esencialmente de todos los enfoques de regresión no-paramétrica, es el hecho de que en la regresión lineal local, generalmente, pueden buscarse modelos en los cuales las funciones de regresión tienen, al menos, la forma:

$$g(s) = \sum_j s_j(x) m_j(x),$$

donde $m_j(x)$ es una función lineal distinta de x y $s_j(x)$ es un peso probabilístico el cual varia a través del espacio designado.

Por supuesto, que los métodos mencionados pueden ser analizados desde el enfoque Bayesiano. Este trabajo, unicamente se restringirá al análisis Bayesiano del modelo de regresión lineal.

3.1. Regresión lineal Bayesiana

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. Las aplicaciones de la regresión son numerosas y empleadas en diversos campos. En efecto, el análisis de regresión puede ser la técnica estadística más ampliamente usada.

Los modelos de regresión son empleados para:

- Descripción de datos.
- Estimación de Parámetros.
- Predicción y estimación.
- Control.

El modelo de regresión lineal simple, es un modelo con un regresor único, x, que tiene una relación con una respuesta, y, la cual se supone es una línea recta. Este modelo es:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{3.1.1}$$

donde β_0 (el intercepto) y β_1 (la pendiente) son constantes desconocidas y ϵ es un error aleatorio. La componente aleatoria se asume que tendra media cero y varianza constante desconocida y, adicionalmente, que son no correlacionadas.

Es conveniente considerar que el regresor x está medido con un error insignificante, mientras que la respuesta y es una variable aleatoria. Esto es, y tiene una distribución de probabilidad en cada posible valor de x. La esperanza de esta distribución es:

$$E[y \mid x] = \beta_0 + \beta_1 x, \tag{3.1.2}$$

y la varianza:

$$V[y \mid x] = V[\beta_0 + \beta_1 x + \epsilon] = \sigma^2. \tag{3.1.3}$$

Observe que la esperanza de y es una función lineal de x pero la varianza de y no depende del valor de x. Además, dado que los errores son no correlacionados, las respuestas también son no correlacionadas.

La pendiente β_1 es el cambio en la esperanza de la distribución de y producida por un cambio unitario en x. Si el rango de los valores sobre x incluyen al cero, entonces la intersección β_0 es el promedio de la distribución de la respuesta y cuando $x = 0$, en caso contrario, β_0 no tiene interpretación práctica.

Los parámetros β_0 y β_1 al ser constantes desconocidas pueden estimarse empleando una muestra de datos. Un criterio empleado para estimar el valor de tales constantes es el método de mínimos cuadrados que no es más que la suma mínima de los cuadrados de las diferencias entre las observaciones y_i y la línea recta. A saber, los estimadores por mínimos cuadrados de la intercepción y la pendiente son:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{3.1.4}$$

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \frac{(\sum y_i)(\sum x_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}, \tag{3.1.5}$$

donde $\overline{y}$ y $\overline{x}$ son las medias aritméticas de y_i y x_i respectivamente. Este método puede ser empleado sin importar la distribución que tengan los errores ϵ. En caso de que sea conocida la distribución de los errores, un método alternativo de estimación, es el método de máxima verosimilitud.

Para el caso de p regresores expresaremos el modelo lineal general en forma matricial, esto es:

$$Y = X\beta + \epsilon, \tag{3.1.6}$$

donde $Y = (Y_1, Y_2, ..., Y_n)'$ es un vector $n \times 1$ de observaciones, $X = [x_{ij}]$ una matriz $n \times p$ de constantes conocidas, β un vector $p \times 1$ de coeficientes de regresión desconocidos y $\epsilon = (\epsilon_1, ..., \epsilon_n)'$ un vector $p \times 1$ de varialbles aleatorias con media cero.

En adelante se supondra que ϵ tiene una distribución normal multivariante, es decir:

$$\epsilon \sim N_n(0, \sigma^2 I_n), \tag{3.1.7}$$

donde I_n denota la matriz identidad de orden n y $p = k + 1$. En otras palabras:

$$Y \sim N_n(y \mid X\beta, \sigma^2 I_n). \tag{3.1.8}$$

Suponiendo, además, que la matriz X es de rango completo p.

La función de verosimilitud para el modelo anterior es de la forma:

$$L(\beta, \sigma^2, y) \propto \frac{1}{2\pi\sigma^2}^{-\frac{n}{2}} exp\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\}, \tag{3.1.9}$$

y los estimadores de máxima verosimilitud para β y σ^2 estan dados por:

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{3.1.10}$$

$$\hat{\sigma^2} = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta}), \tag{3.1.11}$$

de donde se sigue que:

- $\hat{\beta}$ es un vector de estadísticas suficientes conjuntas para β[1].
- $\hat{\beta}$ y $\hat{\sigma^2}$ son suficientes cojuntamente para (β, σ^2).
- $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$[2].

Desde la perspectiva Bayesiana [15, 23, 33], y considerando la anterior formulación general del modelo de regresión, la función de verosumilitud puede escribirse como:

$$L(\beta, \sigma^2, y) \propto (\sigma^2)^{-\frac{n}{2}} exp\{-\frac{1}{2\sigma^2}[(\beta - \hat{\beta})X'X(\beta - \hat{\beta}) + n\hat{\sigma^2}]\}. \tag{3.1.12}$$

Dada la forma de la verosimilitud (3.1.12), una familia conjugada particularmente conveniente tiene una densidad de la forma[3]:

$$\pi(\beta, \tau) \propto \tau^{\frac{n_0}{2}} exp\{-\frac{\tau}{2}[(\beta - b_0)'B_0(\beta - b_0) + s_0]\}, \tag{3.1.13}$$

[1] Si, σ^2 es conocida.
[2] Distribución Normal multivariante.
[3] En adelante consideraremos $\tau = \frac{1}{\sigma^2}$.

donde $n_0, s_0 \in \Re$ y $b_0 \in \Re^p$, B_0 matriz $p \times p$ simétrica y semidefinida positiva. Note que dado el valor de τ, el kernel de la densidad condicional $\pi(\beta \mid \tau)$ es proporcional a la densidad $N_p(\beta \mid b_0, \tau^{-1}B_0^{-1})$, es decir:

$$\pi(\beta \mid \tau) \propto \tau^{\frac{p}{2}} exp\{-\frac{\tau}{2}(\beta - b_0)'B_0(\beta - b_0)\}. \tag{3.1.14}$$

De manera que los factores restantes corresponden a la densidad marginal de $\pi(\tau)$, es decir:

$$\pi(\tau) \propto \tau^{\frac{n_0-p}{2}} exp\{\frac{-s_0\tau}{2}\}. \tag{3.1.15}$$

Haciendo $a = n_0 - p + 2$ y $d = s_0$, se tiene entonces, que:

$$\pi(\beta, \tau) = \pi(\beta \mid \tau)\pi(\tau) = N_p(\beta \mid b_0, \tau^{-1}B_0^{-1})Ga(\tau \mid \frac{a}{2}, \frac{d}{2}). \tag{3.1.16}$$

Esta distribución es conocida como Normal-Gamma y es propia si $a > 0$, $d > 0$ y B_0 es definida positiva.

En situaciones en las que se desea representar un estado de información iniclal vaga acerca de (β, τ), es común utilizar algún tipo de distribución inicial *no informativa*. Uno de los métodos más conocidos para obtener dichas distribuciones es la Regla de Jeffreys. Así, en el caso de la regresión que se esta considerando, la distribución inicial de Jeffreys es:

$$\pi(\beta, \tau) \propto \tau^{\frac{(p-2)}{2}}. \tag{3.1.17}$$

Esta distribución es impropia y puede ser obtenida de la familia conjugada anterior si $a = d = 0$ y $B_0 = 0$.

Otra distribución no informativa comunmente empleada para los parámetros de localización y escala es $\pi(\beta, \tau) \propto \tau^{-1}$, la cual corresponde a la distribución inicial de referencia obtenida a partir del método expuesto en [4]. Al igual que la distribución de Jeffreys, es imporpia y es un caso límite de la familia conjuda cuando $a = -p$, $d = 0$ y $B_0 = 0$.

En ambos casos, B_0 implica que la varianza de la distribución inicial de β es infinita, lo que generalmente se interpreta como una forma de representar un estado de información inicial vaga acerca del valor de β.

Puede demostrarse que la distribución final de (β, τ) para el modelo de regresión considerado, teniendo en cuenta que se esta empleando una distribución inicial conjugada, es de la forma:

$$\pi(\beta, \tau \mid y) = N_p(\beta \mid b_1, \tau^{-1}B_1^{-1})Ga(\tau \mid \frac{a_1}{2}, \frac{d_1}{2}), \tag{3.1.18}$$

donde:

$$b_1 = (X'X + B_0)^{-1}(Xy + B_0b_0),\ B_1 = X'X + B_0,$$

$$a_1 = n + a$$

$$d_1 = (y - Xb_1)'(y - Xb_1) + (b_1 - b_0)'B_0(b_1 - b_0) + d.$$

Así, la distribución final de τ es $\pi(\tau \mid y) = Ga(\tau \mid \frac{a_1}{2}, \frac{d_1}{2})$, lo que implica que la distribución final para σ^2 es $IGa(\sigma^2 \mid \frac{a_1}{2}, \frac{d_1}{2})$.

Por otro lado, para hacer inferencias sobre β, es necesario calcular su distribución marginal final, dada por:

$$\pi(\beta \mid y) = \int \pi(\beta, \tau \mid y)d\tau, \tag{3.1.19}$$

es decir,

$$\pi(\beta \mid y) = \frac{\Gamma(\frac{a_1+p}{2})}{\Gamma(\frac{a_1}{2})\pi^{\frac{p}{2}}} det\{\frac{1}{d_1}B_1\}^{\frac{1}{2}}\{1 + \frac{1}{d_1}(\beta - b_1)'B_1(\beta - b_1)\}^{\frac{-(a_1+p)}{2}}. \tag{3.1.20}$$

En otras palabras, $\pi(\beta \mid y) = S_{t_p}(\beta \mid b_1, T_1^1, a_1)$, donde $T_1 = (\frac{a_1}{d_1})B_1$, de manera que la distribución final de β es una t de Student multivariante con a_1 grados de libertad.

3.2. Aplicación: Predicción de dióxido de carbono en Veracruz

Los estudios sobre el dióxido de carbono (uno de los gases causantes del efecto invernadero) son de suma importancia debido a su relación con el fenómeno del calentamiento global. Estos gases son continuamente emitidos y removidos en la atmósfera por procesos naturales sobre la tierra, pero las actividades antropogénicas causan cantidades adicionales de los mismos, incrementando sus concentraciones en la atmósfera, lo que tiende a sobrecalentarla.

En la actualidad, existe gran interés científico en modelar la relación que guardan las emisiones del dióxido de carbono (CO_2) con las propiedades del suelo y, con el propósito de estudiar tales asociaciones, en esta sección se presenta el ajuste Bayesiano como una aplicación específica a la teoria presentada a lo largo de este trabajo de tesis.

Con el proposito de modelar las emisiones de dióxido de carbono con diferentes propiedades del suelo se realizaron, en determinados sitios de muestreo del Parque Ecológico Jaguaroundi en Coatzacoalcos, Veracruz, México [36]; mediciones a través del método de cámara estática, colocando 3 cámaras en los sitios seleccionados, buscando homogeneidad en el terreno y considerando un mismo plano por altitud así como cubierta vegetal. Las emisiones registradas fueron analizadas con un Detector de Ionización de Flama y las propiedades físicas y químicas de las muestras de suelo fueron analizadas por el Grupo de Edafología del Instituto de Geología de la UNAM.

Los 24 datos muestrales a analizar estan dados en la Tabla 3.1:

Datos					
CO_2	321.890	283.620	425.380	-228.000	531.500
	273.060	10.526	9.487	-23.339	36.291
	-8.039	25.954	304.231	100.039	130.866
	59.759	126.388	99.571	44.376	101.781
	53.160	148.698	-15.208	156.298	
Nitrógeno	0.675	0.480	1.130	0.170	0.265
	0.230	0.115	0.240	0.285	0.245
	0.190	0.255	0.185	0.220	0.160
	0.280	0.260	0.280	0.105	0.150
	0.165	0.135	0.105	0.135	

Tabla 3.1: Mediciones de emisiones de CO_2 y Nitrógeno

Ajuste Bayesiano

El modelo de regresión lineal a considerar, para la modelación de la relación que guardan las emisiones de dióxido de carbono con las propiedades del suelo, es el dado en (3.1.8), es decir:

$$Y \sim N_n(y \mid X\beta, \sigma^2 I_n),$$

cuya componente sistemática es:

$$\mu = X\beta,$$

El modelo de regresión lineal Bayesiano se caracteriza por precisar una distribución inicial para los parámetros del modelo. Esta distribución, debe describir la incertidumbre sobre los verdaderos valores de los parámetros y, una vez observados los datos, la información sobre los parámetros estará contenida en la distribución final la cual es deducida a través del Teorema de Bayes estableciendo, ésta distribución, la solución Bayesiana al problema de inferencia planteado.

De acuerdo a lo presentado en la sección 2.3, el proceso inferencial Bayesiano sobre una magnitud de interés relacionada a un conjunto de observaciones demanda:

1. Especificar un *modelo probabilístico* -$p(CO_2 \mid \beta, \sigma^2)$.

   - El modelo probabilístico a considerar, para ajustar la relación que guardan las emisiones de dióxido de carbono (CO_2) con las propiedades del suelo, en particular con el Nitrógeno (Ni), es el dado en (3.1.12):

$$p(CO_2 \mid \beta, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} exp\{-\tfrac{1}{2\sigma^2}[(\beta - \hat{\beta})N'N(\beta - \hat{\beta}) + n\hat{\sigma^2}]\}^4.$$

2. Especificar una *diststribución inicial* -$\pi(\beta, \sigma^2)$.

   - Para los parámetros desconocidos (β y σ^2) del modelo probabilistico a considerar, debe especificarse una distribución inicial conjunta que describa la incertidumbre que se tiene sobre sus verdaderos valores.

     Las distribuciones iniciales a considerar son[5]:

     - Informativa:
       - Conjugada:
         $\pi_1(\beta, \tau) = \pi(\beta \mid \tau)\pi(\tau)^6 = N(\beta \mid b_0, \tau^{-1}B_0)Ga(\tau \mid \frac{a}{2}, \frac{d}{2})$.
       - Semi-conjugada:
         $\pi_2(\beta, \tau) = \pi(\beta)\pi(\tau)^7 = N(\beta \mid b_0, B_0^{-1})Ga(\tau \mid \frac{a}{2}, \frac{d}{2})$.
     - No-Informativa.
       - Inicial de Jeffreys:
         $\pi_3(\beta, \tau)^8 \propto \tau$.

3. Determinar la *distribución final* -$\pi(\beta, \sigma^2 \mid CO_2)$.

   - La determinación de la distribución final, solución Bayesiana a un problema de inferencia, se obtiene a través del Teorema de Bayes:

$$\pi(\beta, \tau \mid CO_2) \propto l(CO_2 \mid \beta, \tau)\pi(\beta, \tau)^9.$$

     En particular, para $\pi_1(\beta, \tau)$ (o bien para $\pi_2(\beta, \tau)$), la distribución final conjunta de los parámetros es la dada en (3.1.18):

$$\pi(\beta, \tau \mid CO_2)^{10} = N_p(\beta \mid b_1, \tau^{-1}B_1^{-1})Ga(\tau \mid \tfrac{a_1}{2}, \tfrac{d_1}{2}).$$

     Especificamente, de acuerdo al modelo de regresión empleado, el interés recae en realizar inferencias sobre β, por lo que es necesario calcular su *distribución marginal final*, la cual está dada por (3.1.20):

$$\pi(\beta \mid CO_2) = \frac{\Gamma(\frac{a_1+p}{2})}{\Gamma(\frac{a_1}{2})\pi^{\frac{p}{2}}} det\{\tfrac{1}{d_1}B_1\}^{\frac{1}{2}}\{1 + \tfrac{1}{d_1}(\beta - b_1)'B_1(\beta - b_1)\}^{\frac{-(a_1+p)}{2}}.$$

     Por lo tanto, la distribución final de β es una t de Student multivariante con a_1 grados de libertad:

$$\pi(\beta \mid CO_2) = S_{t_p}(\beta \mid b_1, T_1^1, a_1) \text{ donde } T_1 = (\tfrac{a_1}{d_1})B_1.$$

[4]Cuando $p(CO_2 \mid \beta, \sigma^2)$ se considerada como una función no de CO_2 pero si de (β, σ^2), es llamada función de *verosimilitud*, la cual denotaremos por: $l(\beta, \sigma^2 \mid CO_2)$.
[5]En adelantes $\tau = \frac{1}{\sigma^2}$.
[6]La especificación analítica de esta distribución esta dada en (3.1.13).
[7]Note que se está suponiendo *a priori*, que los parámetros son independientes.
[8]Esta distribución inicial se obtiene de la incial conjugada tomando $a = d = 0$ y $B_0 = 0$.
[9]En otras palabras, el Teorema de Bayes afirma: $Posterior \propto Verosimilitud \times Inicial$.
[10]Considerando $\pi_3(\beta, \tau)$, $\pi(\beta, \tau \mid CO_2)$ tiene una forma analítica no conocida.

El empleo de las técnicas Monte Carlo basadas en cadenas de Markov (MCMC), a través de un software de cómputo, además, de permitir el estudo de modelos Bayesianos más complejos, facilitan de manera considerable el cálculo de la distribución final, así como el de ciertas características de interés (media, mediana, moda) de la distribución final.

Con el propósito de mostrar las técnicas MCMC en la modelación descrita anteriormente, a continuación, se emplearán los siguientes softwares estadísticos Bayesianos:

- WinBUGS.
- R (aplicando el paquete):
  - MCMCpack.
- WinBUGS/BOA[11].

WinBUGS

- Codificación del Modelo:
  - $CO_2 \sim N(CO_2 \mid \mu, \tau)$ donde $\mu = \beta_0 + \beta_1 Ni$.
  - $\beta = (\beta_0, \beta_1) \sim N(\beta \mid b_0, \tau^{-1} B_0^{-1})$.
  - $\tau \sim Ga(\tau \mid \frac{a}{2}, \frac{d}{2})$.
  - $\sigma^2 = \frac{1}{\tau}$.

  es decir, se emplea la distribución incial:

$$\pi_1(\beta, \tau) = \pi(\beta \mid \tau)\pi(\tau) = N(\beta \mid b_0, \tau^{-1} B_0) Ga(\tau \mid \tfrac{a}{2}, \tfrac{d}{2}).$$

- Especificaciones:
  - Valores de los hiperparámetros:
    - Parámetros para la distribución de $\beta = (\beta_0, \beta_1)$:
      - Media inicial: $b_0 = 0$,
      - Precisión incial: $\tau^{-1} B_0^{-1} = 0{,}00001$.
    - Parámetros para la distribución de τ:
      - Parámetro de forma: $\frac{a}{2} = 0{,}001$, .
      - Parámetro de escala: $\frac{d}{2} = 0{,}001$.

    Bajo éstos valores para los hiperparámetros, la distribución conjugada está representando un estado de información inicial *mínimo* informativo sobres (β, τ).
  - Cadenas de valores muestrales generadas: 2.

    WinBUGS ofrece la posibilidad de generar varias cadenas por parámetro, siendo esto una estrategia para mejorar los cálculos y acortar el tiempo de convergencia a la distribución estacionaria.
  - Valores iniciales por cadena:
    - $\beta_0 = 0$, $\beta_1 = 0$, $\tau = 1$.
    - $\beta_0 = 15$, $\beta_1 = 403$, $\tau = 5$[12].

[11] Las cadenas han sido generadas con WinBUGS y el análisis de estas se ha efectudo en R, empleando el paquete BOA (Bayesian Output Analysis) .

[12] Este conjunto de valores iniciales, se han propuesto de acuerdo a la estimación obtenida con el paquete Zelig (Ver ápendice B).

- Iteraciones realizadas por cadena: 10000.

  Se consideraron 8000 valores muestrales por cadena, descartando las primeras 4000 iteraciones en cada cadena, debido a que se encontraban en un proceso de estabilización.

  - La Figura 3.1 muestra la estabilización y convergencia de las cadenas en cada uno de los parámetros del modelo.

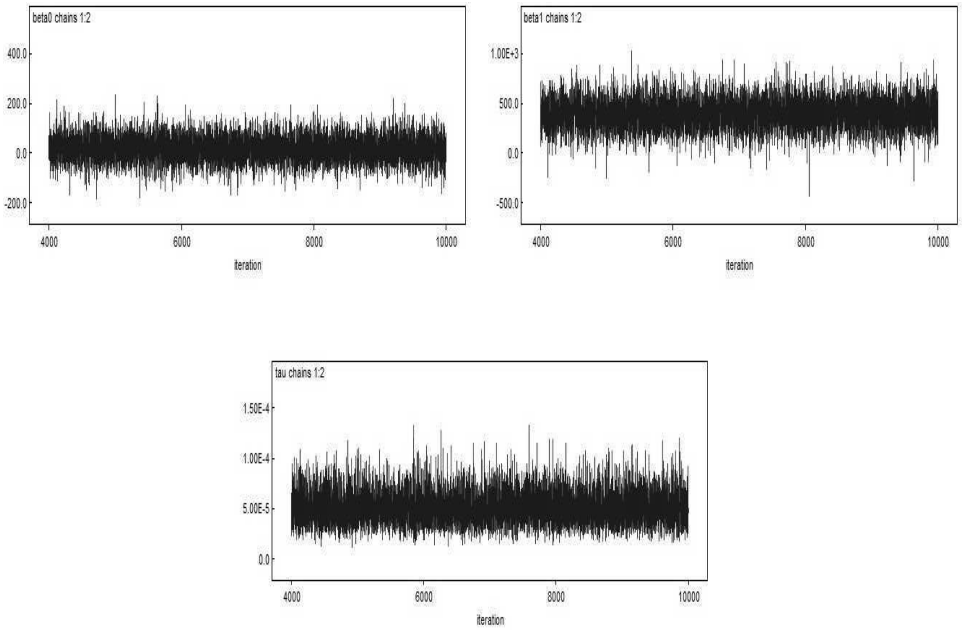


Figura 3.1: Simulación

  - Un criterio para confirmar la convergencia de una cadena a su distribución estacionaria, es el estadístico de Gelman-Rubin. WinBUGS sólo presenta el gráfico del mismo (Figura 3.2), corroborandose, al menos gráficamente, que las cadenas por parámetro convergen.

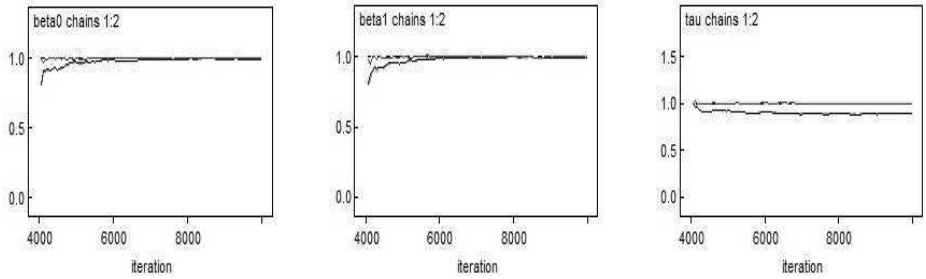


Figura 3.2: Estadístico Gelman-Rubin

  - La Figura 3.3, muestra la autocorrelación entre los valores simulados por cadena para cada parámetro. Aunque WinBUGS sólo muestra un rezago (lag) de 50, puede considerarse que los valores por cadena, para cada parámetro, son independientes.

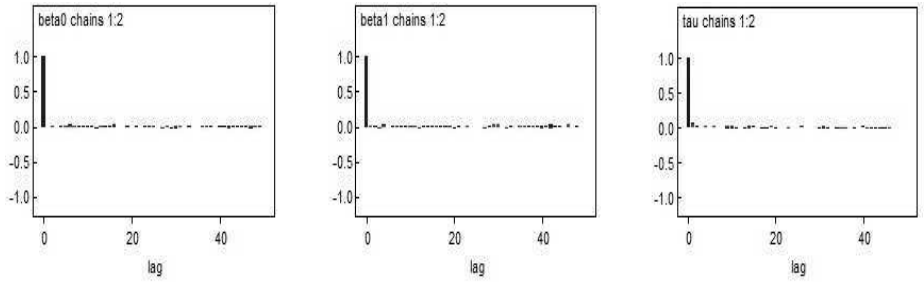


Figura 3.3: Autocorrelaciones

- Inferencias:

  La distribución final así como el resumen estadístico proporcionado por WinBUGS, resulta de la combinación de las cadenas generadas por parámetro.

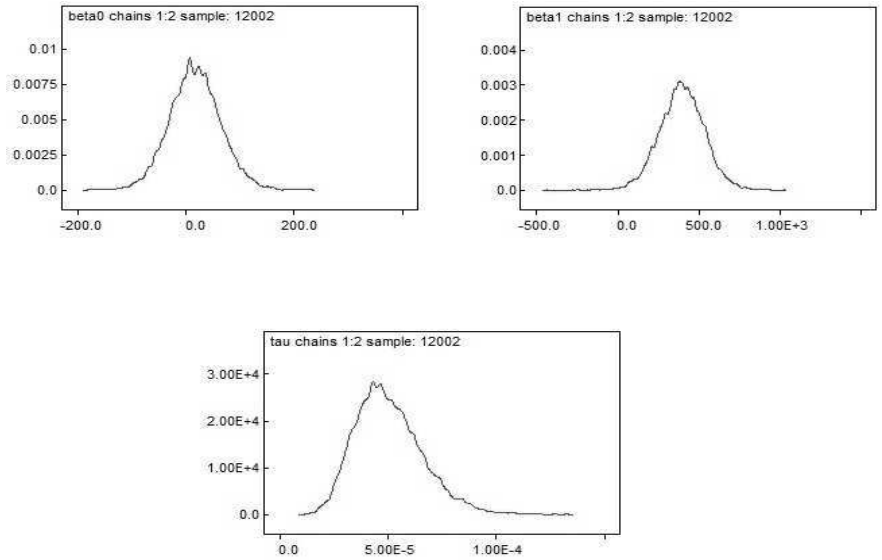


Figura 3.4: Distribución final- por parámetro

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	17.14	47.25	0.4479	-74.05	16.66	112.1	4000	12002
beta1	395.3	138.2	1.345	119.3	395.8	666.8	4000	12002
tau	5.047E-5	1.543E-5	1.501E-7	2.524E-5	4.869E-5	8.526E-5	4000	12002

Figura 3.5: Resumen analítico y estimaciones

MCMCpack

- Codificación del Modelo:
  - $CO_2 \sim N(CO_2 \mid \mu, \tau)$ donde $\mu = \beta_0 + \beta_1 Ni + \epsilon$, con $\epsilon \sim N(\epsilon \mid 0, \tau^{-1})$.
  - $\beta = (\beta_0, \beta_1) \sim N(\beta \mid b_0, B_0^{-1})$.
  - $\tau \sim Ga(\tau \mid \frac{a}{2}, \frac{d}{2})$.

  es decir, se empleará la distribución incial:

$$\pi_2(\beta, \tau) = N(\beta \mid b_0, B_0^{-1}) Ga(\tau \mid \tfrac{a}{2}, \tfrac{d}{2}).$$

- Especificaciones:
  - Valores de los hiperparámetros:
    Se consideraron por defecto, los valores definidos en la función MCMCregress, es decir:
    - Media incial para β: $b_0 = 0$.
    - Precisión incial para β: $B_0^{-1} = 0$.
    - Parámetros para la distribución de τ:
      - Parámetro de forma: 0,001,
      - Parámetro de escala: 0,001.

    La distribución semi-conjugada bajo los valores específicados para los hiperparámetros, está representando un estado de información inicial *poco* informativo sobres (β, τ).
  - Cadenas de valores muestrales generadas: 1.
    MCMCpack, a través de la función *MCMCregress*, genera sólo una cadena MCMC.
  - Iteraciones realizadas: 3771.
    MCMCpack, a través del paquete CODA, ofrece un diagnostico de control[13] que calcula el número de iteraciones necesarias a realizar, ofeciendo también, diagnósticos de convergencia.
    - De acuerdo al diagnostico de control (Figura 3.6), el número necesario de iteraciones a realizar para estimar el cuantil posterior $q = 0{,}025$ con tolerancia $\pm r = 0{,}005$ y probabilidad $s = 0{,}95$ de la distribución final por parámetro, debe ser de $N = 3771$, sin considerar las primeras $M = 2$ iteraciones en cada parámetro.

```
> MCMCchain = MCMCregress(CO2 ~ NI, data = data)
> raftery.diag(MCMCchain, q=0.025, r=0.005, s=0.95, converge.eps=0.001)

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

             Burn-in  Total Lower bound  Dependence
             (M)      (N)   (Nmin)       factor (I)
 (Intercept) 2        3771  3746         1.01
 NI          2        3771  3746         1.01
 sigma2      2        3710  3746         0.99
```

Figura 3.6: Raftery and Lewis's diagnostic

[13]Diagnóstico basado sobre un criterio de precisión en la estimación de algún cuantil posterior q, con tolerancia $\pm r$ y probabilidad s.

- La tabla (Figura 3.7) detalla la autocorrelación entre los valores simulados en cada parámetro en diferentes rezagos (lag), evidenciando la independencia entre tales valores, gráficamente (Figura 3.8) sólo se muestra la autocorrelación para un rezago (lag) de 30.

```
AUTOCORRELATIONS WITHIN EACH CHAIN:
===================================

Iterations used = 3:3773
Thinning interval = 1
Sample size per chain = 3771

$chain1
, , (Intercept)

        (Intercept)        NI   sigma2
Lag 0      1.000000 -0.77959  0.01793
Lag 1     -0.003853  0.01511  0.00801
Lag 5      0.000573  0.00191 -0.01836
Lag 10    -0.013374  0.00755  0.02518
Lag 50    -0.038564  0.05427  0.01504

, , NI

        (Intercept)        NI   sigma2
Lag 0     -0.779585  1.00000 -0.0213
Lag 1      0.027099 -0.03338 -0.0213
Lag 5     -0.000995 -0.00364  0.0158
Lag 10    -0.001126  0.01157 -0.0235
Lag 50     0.034607 -0.04722 -0.0209

, , sigma2

        (Intercept)        NI   sigma2
Lag 0       0.01793 -0.02127  1.00000
Lag 1       0.00468 -0.01389  0.07720
Lag 5       0.01657  0.00173 -0.01190
Lag 10     -0.01337  0.00376  0.03893
Lag 50     -0.03697  0.01866 -0.00554
```

Figura 3.7: Autocorrelaciones

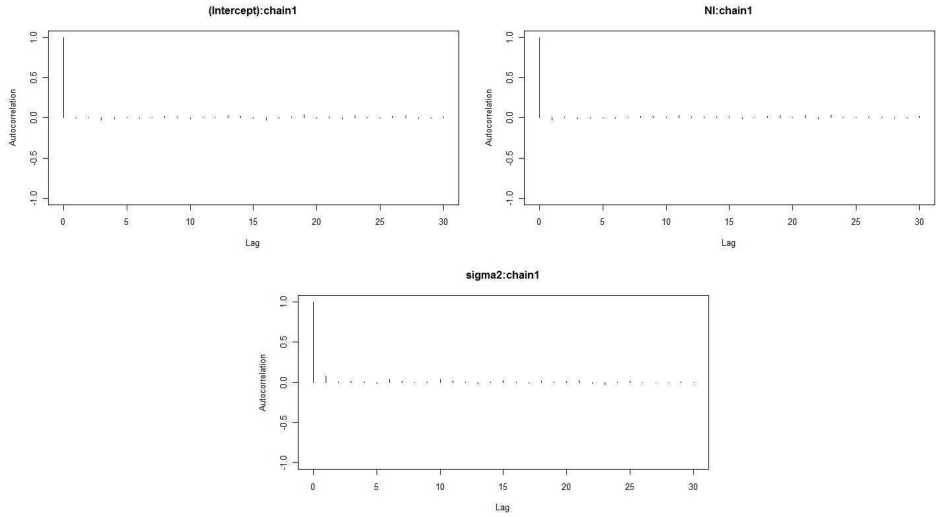
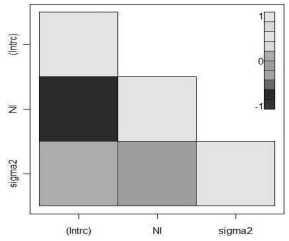


Figura 3.8: Gráfico Autocorrelaciones

- La matriz de correlación entre parámetros (Figura3.9) revela que la convergencia, de las cadenas de los parámetros β_0 y β_1, se da de manera lenta, debido a que la correlación entres estos parámetros es relativamente alta.




```
$chain1
            (Intercept)      NI  sigma2
(Intercept)      1.0000 -0.7796  0.0179
NI              -0.7796  1.0000 -0.0213
sigma2           0.0179 -0.0213  1.0000
```

Figura 3.9: Matriz de correlación

- Inferencias:
  - Los resultados analíticos y diferentes cuantiles para los parámetros de interés, se encuentran contenidos en la Figura 3.10:

```
chain1

Iterations = 3:3773
Thinning interval = 1
Number of chains = 1
Sample size per chain = 3771

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                 Mean       SD Naive SE Time-series SE
(Intercept)     15.72    48.07   0.7828         0.8295
NI             403.07   137.98   2.2470         2.7202
sigma2       21780.16  7141.45 116.2943       151.4662

2. Quantiles for each variable:

                 2.5%      25%      50%      75%   97.5%
(Intercept)    -80.83   -15.39    16.07    47.53   110.7
NI             124.18   314.56   401.80   494.34   677.4
sigma2       11871.22 16872.59 20486.80 25095.49 39659.6
```

Figura 3.10: Resumen Analítico

- La representación gráfica de la distribución final así como la traza de las simulaciones realizadas por parámetro, se encuentra dada en la Figura 3.11:

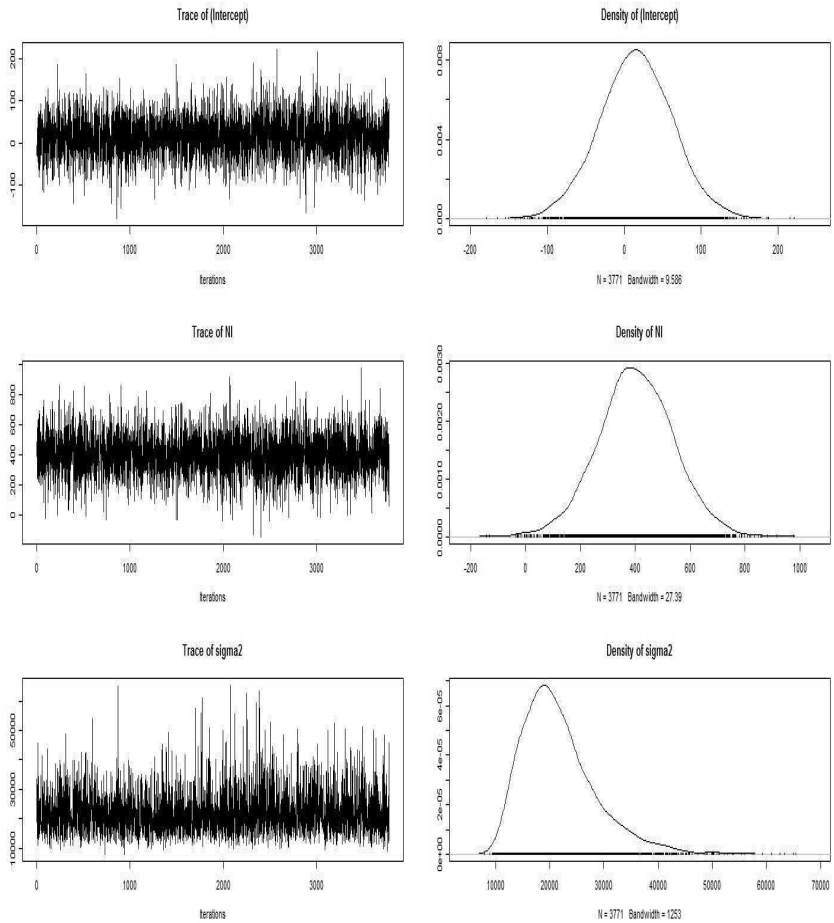


Figura 3.11: Traza y Distribución Final- por parámetro

WinBUGS-BOA

- Codificación del modelo:
  - $CO_2 \sim N(CO_2 \mid \mu, \tau)$ donde $\mu = \beta_0 + \beta_1 \cdot Ni$.
  - $\beta_0, \beta_1 \sim N(\beta_i \mid 0, 0{,}00001)$ para $i = 0, 1$.
  - $\sigma \sim Un(\sigma \mid 0, 1000)$.
  - $\tau = \frac{1}{\sigma^2}$.

  es decir, se empleará la distribución incial (*No-Informativa*):

$$\pi_3(\beta, \tau) \propto \tau.$$

- Especificaciones:
  - Cadenas de valores muestrales generadas: 2.

    Nuevamente se emplea la estategia de generar varias cadenas, a pesar de que las cadenas fueron generadas en WinBUGS, su análisis se efectuó con el paquete BOA.
  - Valores inciales para los parámetros de interés:
    - $\beta_0 = 0$, $\beta_1 = 0$, $\sigma = 1$.
    - $\beta_0 = 15$, $\beta_1 = 403$, $\sigma = 140$[14].

- Iteraciones realizadas: 5000.

  El número de iteraciones a considerar, se determinó de acuerdo al análisis con BOA, para ello se efectúo un análisis de convergencia por cadena sobre (β_0, β_1, σ).

  - Al simular 3000 valores[15], el diagnóstico de *Raftery and Lewis*; diagnóstico que analiza la convergencia a la distribución estacionaria, estimando el número de iteraciones para determinar el cuantil posterior $q(0{,}025)$ con exactitud $r(\pm 0{,}005)$ y probabilidad $s(0{,}95)$ [38]; estimó que es necesario simular por lo menos 3746 valores.

    Una vez realizadas las simulaciones necesarias, el análisis es categorizado en dos grupos:

    1. Análisis de Convergencia.
    2. Resumen Estadístico.
  - Analisis de Convergencia:
    - La inspección de las autocorrelaciones entre los valores simulados por cadena (para diferentes rezagos), es una manera de analizár su convergencia; autocorrelaciones altas indican una mezcla lenta en las cadenas y, por lo general, una convergencia paulatina a su distribución estacionaria.

```
LAGS AND AUTOCORRELATIONS:
==========================

Chain: cadena1
--------------

            Lag 1         Lag 5       Lag 10       Lag 50
beta0 0.007054019 -0.0001742516  0.001570944 -0.007276596
beta1 0.013998296  0.0142623790 -0.003874971 -0.007191531
sigma 0.157547620 -0.0168942658 -0.016505669 -0.002807125

Press <ENTER> to continue

Chain: cadena2
--------------

            Lag 1        Lag 5       Lag 10      Lag 50
beta0 -0.01411949  0.008958893 -0.007067073  0.04134567
beta1 -0.01461589 -0.010657307 -0.001390570  0.03628924
sigma  0.16914115  0.028940756 -0.001874079 -0.01354043
```

Figura 3.12: Autocorrelaciones por cadena

[14]Este conjunto de valores iniciales se han propuesto de acuerdo a la estimación obtenida con SPSS.

[15]No se consideran los primeros 500 valores, ya que el proceso optimiza el cálculo de las variables estocásticas durante esta etapa de adaptación, descartando tales valores en el análisis posterior.

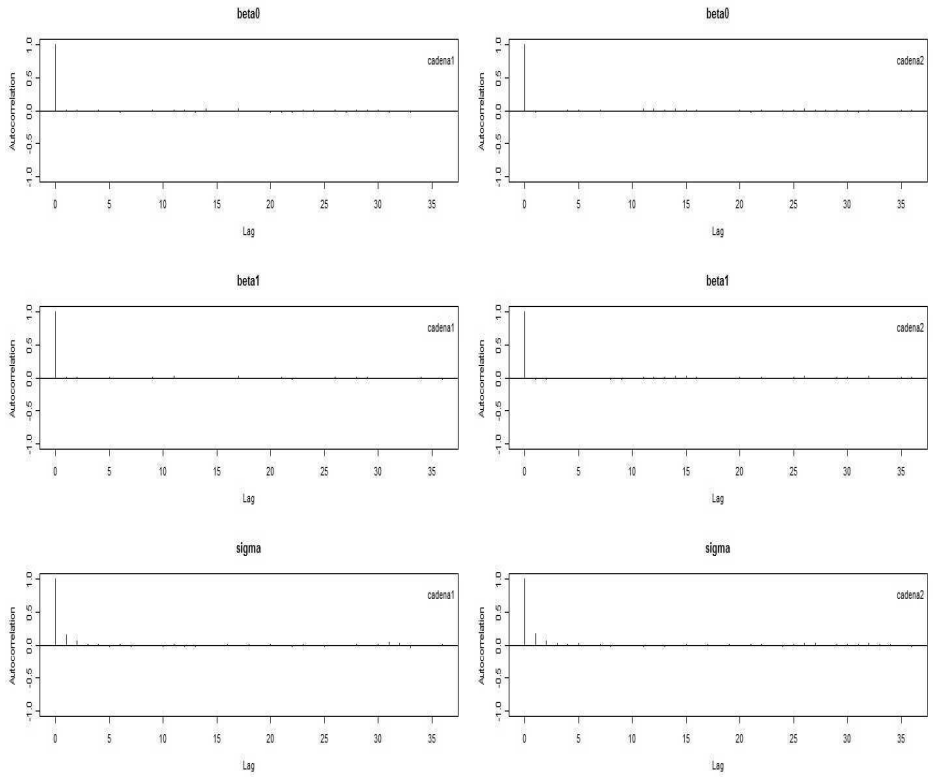


Figura 3.13: Autocorrelaciones

- La correlación entre parámetros, también permite analizar la convergencia de las cadenas; correlaciones altas entre los parámetros, conduce a una convergencia lenta.

```
CROSS-CORRELATION MATRIX:
=========================

Chain: cadena1
--------------

      beta0      beta1      sigma
beta0 1
beta1 -0.7736527 1
sigma 0.017049   -0.0222862 1

Press <ENTER> to continue

Chain: cadena2
--------------

      beta0      beta1     sigma
beta0 1
beta1 -0.7809294 1
sigma 0.0187334  -0.026819 1
```

Figura 3.14: Matriz de Correlaciones

- Una opción gráfica para un análisis intuitivo de convergencia, es el gáfico de trazas[16]; si una cadena ha convergido, sus valores ocilarán al rededor de la moda de la distribución.

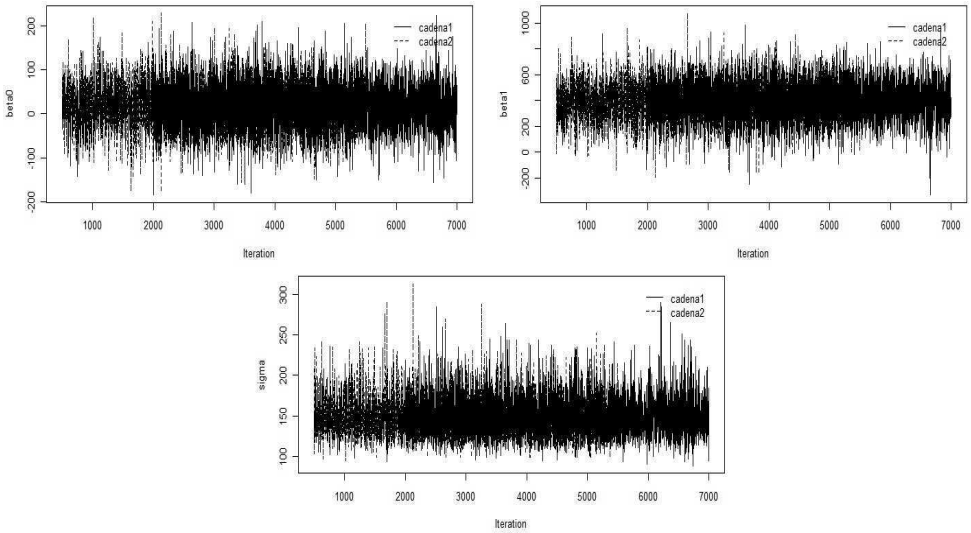


Figura 3.15: Trazas

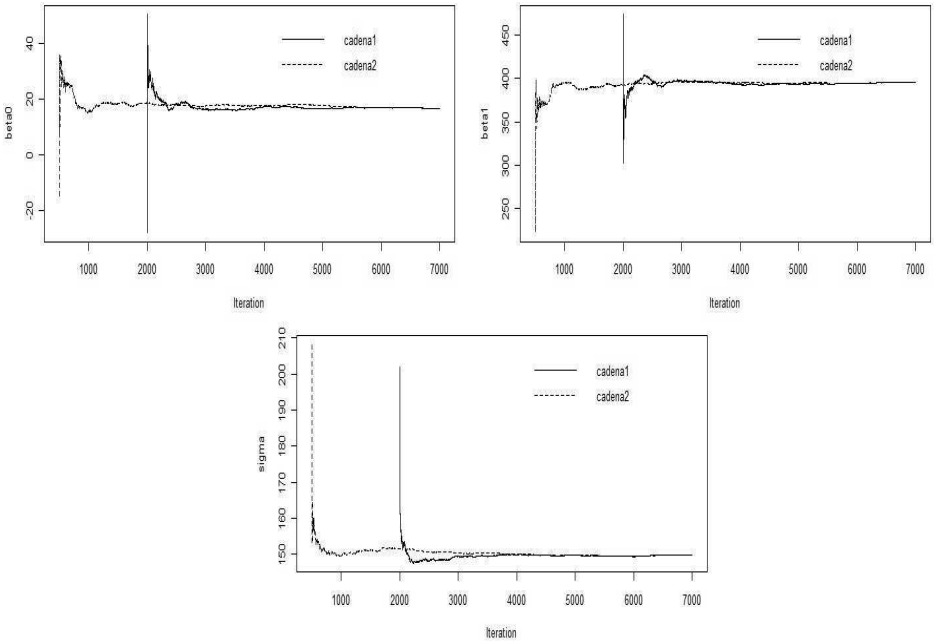


Figura 3.16: Sampler Running Mean

[16]Representación gráfica del número de iteración contra el valor del parámetro.

- *Raftery and Lewis Convergence Diagnostics*, además de indicar, en un principio, el número necesario de simulaciones a realizar, ahora también, inidca el factor de dependencia entre cada valor simulado (Fig. 3.17).

```
RAFTERY AND LEWIS CONVERGENCE DIAGNOSTIC:
=========================================

Quantile = 0.025
Accuracy = +/- 0.005
Probability = 0.95

Chain: cadena1
--------------

      Thin Burn-in Total Lower Bound Dependence Factor
beta0    1       2  4020        3746          1.073145
beta1    1       2  3891        3746          1.038708
sigma    2       4  8628        3746          2.303257

Press <ENTER> to continue

Chain: cadena2
--------------

      Thin Burn-in Total Lower Bound Dependence Factor
beta0    1       2  3891        3746          1.038708
beta1    1       2  3706        3746          0.989322
sigma    1       2  4020        3746          1.073145
```

Figura 3.17: Raftery and Lewis Convergence Diagnostic

- El diagnóstico de *Geweke*, compara la media de la primera parte (10 %) de los valores muestrales generados contra la media de los últimos valores (50 %), a través de un estadístico Z[17] que a medida que el número de simulaciones tiende a infinito, Z tiene a $N(0,1)$[18], por lo tanto, si $p - value < 0{,}05$, existe evidencia contra la no-convergencia[19] [38].

```
GEWEKE CONVERGENCE DIAGNOSTIC:
==============================

Fraction in first window = 0.1
Fraction in last window = 0.5

Chain: cadena1
--------------

             beta0      beta1      sigma
Z-Score 0.9764444 -0.1570977 -1.6424133
p-value 0.3288443  0.8751678  0.1005044

Press <ENTER> to continue

Chain: cadena2
--------------

              beta0      beta1     sigma
Z-Score -0.5599495 0.03192065 0.7353976
p-value  0.5755139 0.97453533 0.4620974
```

[17]Razón de la diferencia entre las medias y el error estándar asintótico de su diferencia, cuya varianza es determinada por la estimación de la densidad espectral. Debe existir un número suficiente de iteraciones entre ambos conjuntos de valores, para asumir que las dos medias son aproximadamente independientes.
[18]Sólo si la cadena ha convergido.
[19]En caso contrario, los resultados no proveerán evidencia alguna contra su convergencia.

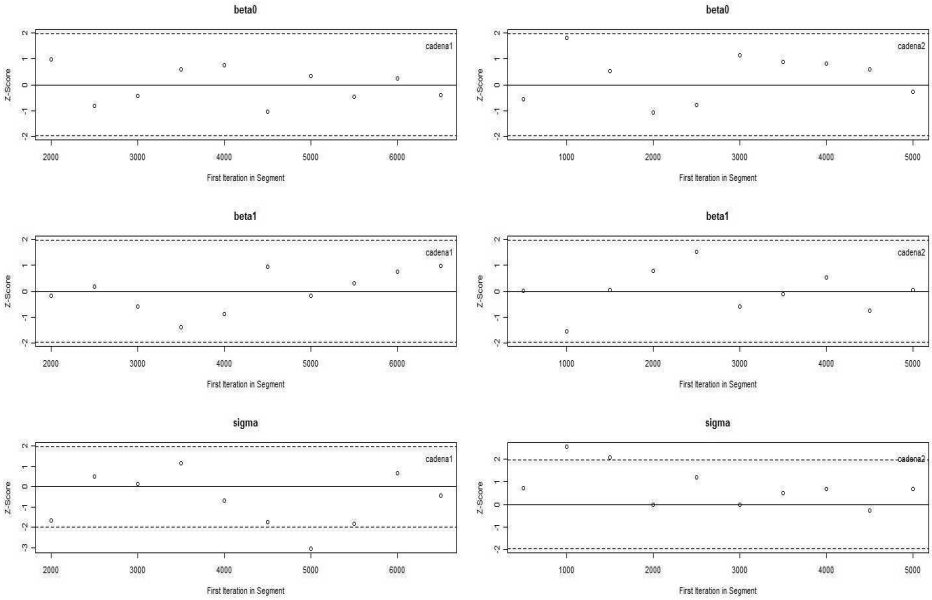


Figura 3.18: Geweke Convergence Diagnostics

- La prueba de *Heidelberger and Welch* [38], otra prueba analítica de convergencia, evidencía estadísticamente si existe no-estacionariedad en los valores simulados. El fallo de esta prueba indica que se deben generar un gran número de valores para logar la convergencia de la cadena. En particular, para las cadenas generadas, no hay evidencia de no-estacionariedad (Fig. 3.19), por lo tanto, cada cadena MCMC converge adecuadamente a su distribución estacionaria.

```
HEIDLEBERGER AND WELCH STATIONARITY AND INTERVAL HALFWIDTH TESTS:
=================================================================

Halfwidth test accuracy = 0.1

Chain: cadena1
--------------

      Stationarity Test Keep Discard    C-von-M Halfwidth Test      Mean Halfwidth
beta0           passed 5001       0 0.08016645         passed  16.71605  1.420293
beta1           passed 5001       0 0.12866439         passed 394.96454  4.210055
sigma           passed 5001       0 0.17748312         passed 149.90996  0.773865

Press <ENTER> to continue

Chain: cadena2
--------------

      Stationarity Test Keep Discard    C-von-M Halfwidth Test      Mean Halfwidth
beta0           passed 5001       0 0.09691092         passed  17.47012  1.595156
beta1           passed 5001       0 0.13376426         passed 395.14306  4.042399
sigma           passed 3501    1500 0.07803704         passed 148.71631  1.014357
```

Figura 3.19: Heidelberger and Welch Convergence Diagnostic

- El diagnóstico *Books, Gelman & Rubin* [38], expone que los valores simulados en cada cadena por parámetro, provienen, en efecto, de su distribución estacionaria.

```
BROOKS, GELMAN, AND RUBIN CONVERGENCE DIAGNOSTICS:
==================================================

Iterations used = 3751:5500

Potential Scale Reduction Factors
---------------------------------

    beta0     beta1     sigma
0.9997938 0.9999125 1.0000385

Multivariate Potential Scale Reduction Factor = 1.000184

Corrected Scale Reduction Factors
---------------------------------

      Estimate    0.975
beta0 0.999916 1.000236
beta1 1.000193 1.000991
sigma 1.000055 1.001359
```

Figura 3.20: Books, Gelman & Rubin Diagnostics Convergence

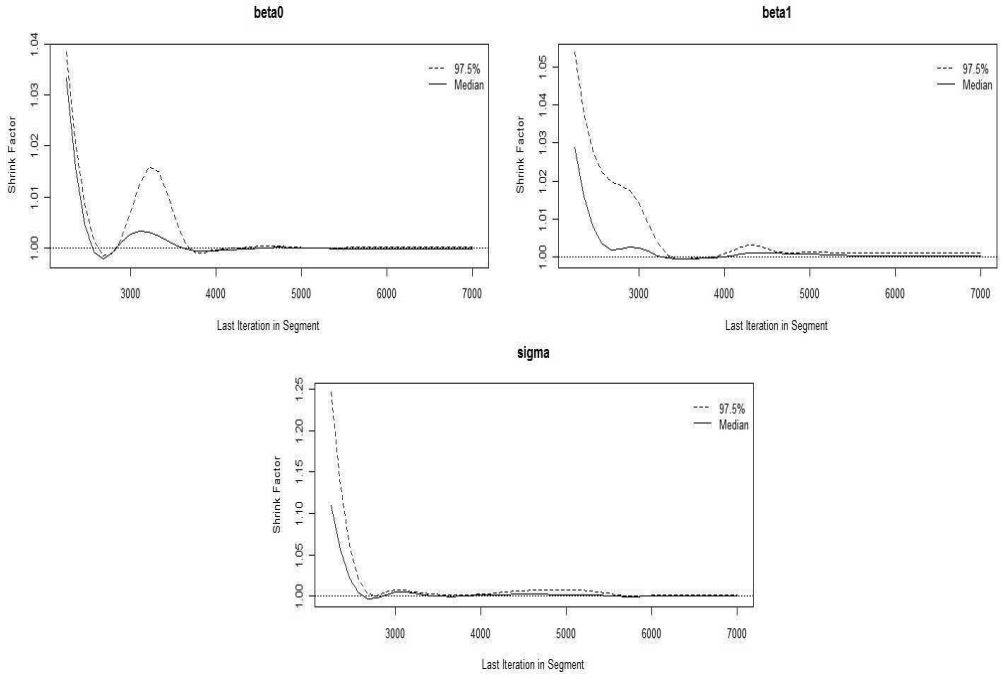
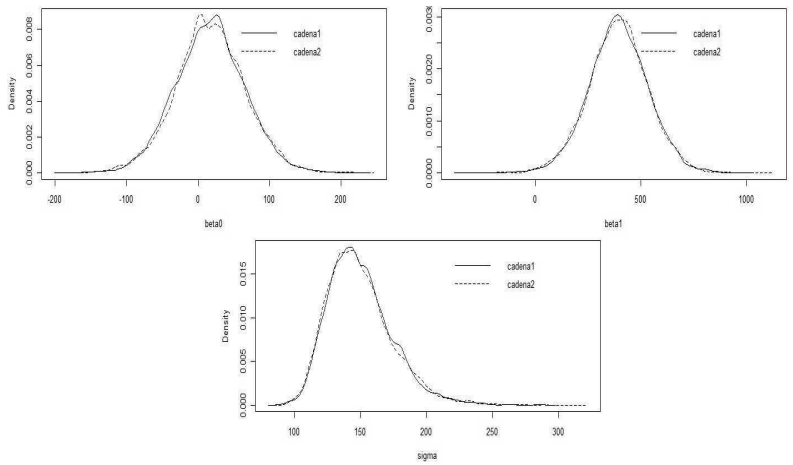


Figura 3.21: Gelman and Rubin Shrink Factors

- Resumen Estadístico:

  - La distribución final estimada y diferentes resultados analíticos por parámetro, en cada cadena, están dados a continuación:




```
SUMMARY STATISTICS:
===================

Bin size for calculating Batch SE and (Lag 1) ACF = 50

Chain: cadena1
--------------

          Mean       SD  Naive SE  MC Error  Batch SE  Batch ACF  0.025    0.5 0.975 MinIter MaxIter Sample
beta0  16.71605  48.86038 0.6909211 0.7246524 0.6905029 0.19385358 -79.05  17.38 112.5    2000    7000   5001
beta1 394.96454 141.97797 2.0076709 2.1480267 2.0458100 0.12011629 113.80 394.20 671.1    2000    7000   5001
sigma 149.90996  24.34638 0.3442754 0.3948363 0.4210102 0.09329152 111.60 146.70 207.0    2000    7000   5001

Press <ENTER> to continue

Chain: cadena2
--------------

          Mean       SD  Naive SE  MC Error  Batch SE  Batch ACF  0.025    0.5 0.975 MinIter MaxIter Sample
beta0  17.47012  49.37749 0.6982333 0.8138699 0.7539459 0.16518758 -80.82  17.17 115.4     500    5500   5001
beta1 395.14306 142.20863 2.0109327 2.0624866 2.0493579 0.07823410 107.70 396.60 675.0     500    5500   5001
sigma 149.58331  24.99884 0.3535016 0.4802423 0.4456429 0.02495561 111.30 146.20 206.4     500    5500   5001
```

Figura 3.22: Resumen Estadístico por cadena

- Inferencias:

  Aunque cualquiera de las dos cadenas generadas permiten producir inferencias, al generar varias cadenas, es común combinarlas para tal fin[20].

[20] EL análisis de autocorrelación y convergencia no deben realizarse a la cadena combinada, ya que potencialmente

```
SUMMARY STATISTICS:
===================

Bin size for calculating Batch SE and (Lag 1) ACF = 50

Chain: cadena1::cadena2
-----------------------

          Mean       SD  Naive SE  MC Error  Batch SE  Batch ACF   0.025    0.5    0.975 MinIter MaxIter Sample
beta0  17.09308  49.11861 0.4911369 0.6083907 0.5431582 0.29607309 -80.059  17.27 114.1950     500    7000  10002
beta1 395.05380 142.08627 1.4207206 1.6465269 1.6637877 0.14267764 111.410 395.50 673.4900     500    7000  10002
sigma 149.74664  24.67407 0.2467160 0.3048385 0.3375272 0.06055621 111.500 146.40 206.6975     500    7000  10002
```

Figura 3.23: Resumen Estadístico

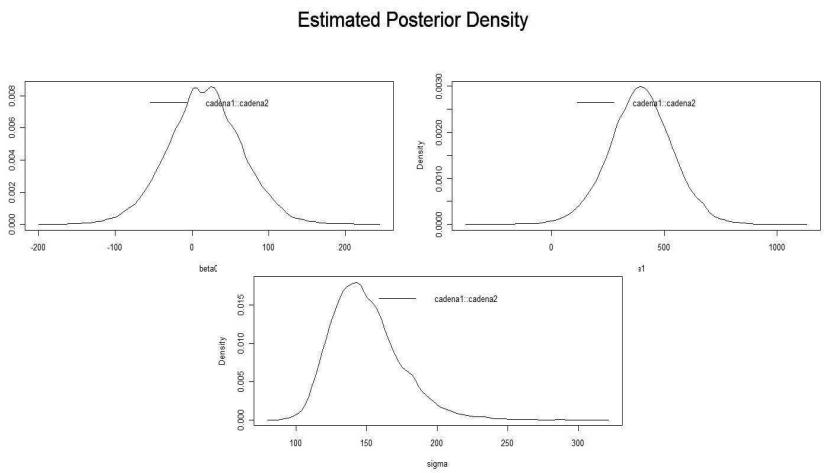


Figura 3.24: Distribución Final

```
HIGHEST PROBABILITY DENSITY INTERVALS:
======================================

Alpha level = 0.05

Chain: cadena1::cadena2
-----------------------

      Lower Bound Upper Bound
beta0      -77.67       116.0
beta1      132.60       691.5
sigma      107.50       198.2
```

Figura 3.25: Highest Probability Density Intervals

3.2.1. Análisis del Modelo de Regresión Bayesiano

La modelación sobre la realción que guardan las emisiones de dióxido de carbono (CO_2) con el nitrogeno (Ni), llevada a cabo con el modelo de regresión lineal Bayesiano, se ha desarrollado reflejando la información incial sobre los parámetros del modelo a través de tres distribuciones inciales. La Tabla 3.2 presenta el resumen de los modelos obtenidos.

Análisis del Modelo de Regresión		
$y = \beta_0 + \beta_1 x + \epsilon$		
Modelo Bayesiano		Modelo No-Bayesiano
Informativa	Conjugada(WinBUGS) $CO_2 = 17.14 + 395.3 \cdot Ni$ $\tau = 5.047E - 5$ $s = 140.76$	(SPSS - Zelig) $CO_2 = 15.12 + 403.31 \cdot Ni$ $s = 140.93$ $\tau = 5.034E - 5$
	Semi-Conjugada (MCMCpack) $CO_2 = 15.75 + 403.07 \cdot Ni$ $\tau = 4.59E - 5$ $s = 147.58$	
No- Informativa	Inicial de Jeffreys (WinBUGS-BOA) $CO_2 = 17.09 + 395.05 \cdot Ni$ $\tau = 4.45E - 5$ $s = 149.74$	
$y = CO_2$, $x = Ni$, $\tau = \frac{1}{\sigma^2}$		

Tabla 3.2: Análisis Bayesiano sobre la predicción del dióxido de carbono.

Como puede observarse, en cuanto a la pendiente (β_1) de los modelos ajustados por la distribución conjugada y la inicial de Jeffreys, no son muy disímiles, sin embargo, existe gran variación entre los estimadores de la presición (τ)[21], el valor más grande[22] del mismo se obtiene cuando se ha utilizado la distribucion inicial conjugada, por lo que puede considerarse que el modelo obtenido, bajo la distribución informativa conjugada, es el mejor. Además, parece describir más realisticamente la relación entre el nitrogeno y el dióxido de carbono, de acuerdo a la información que se tiene sobre los parámetros.

contiene muestras múltiples por iteración.
[21]Equivalentemente, entre los estimadores de la varianza de los errores del modelo (s^2).
[22]Si se emplea s, debe considerarse el valor más pequeño.

Conclusiones

El *Teorema de Bayes* (1763) ha dado origen a lo que hoy se conoce como el paradigma Bayesiano. El comienzo del pensamiento Bayesiano surge a partir del siglo XIX con *Laplace* y, a pesar de que en la primera mitad del siglo XX se desarrollo la estadística frecuentista, el enfoque Bayesiano progresó gracias a la contribución matemática de *Frank Ramsey*, *Bruno de Finetti*, *Harold Jeffreys* y, más recientemente, a *Leonard J. Savage*, *Dennis V. Lindley*, *José M. Bernardo* entre otros. Pero es hasta finales del siglo XX, una vez desarrollados los ordenadores y los métodos computacionales, que la inferncia Bayesiana ha comenzado a emplearse, llevando no sólo a la investigación en su metodología, sino, al uso de sus métodos en diversas áreas de aplicación debido, principalmente, a su coherencia filosófica y a la proyección de los problemas estadísticos dentro del marco de la toma de decisiones.

Consideramos que este trabajo modestamente ha presentado, como uno de sus principales objetivos, una introducción a la metodología Bayesana a través de los siguientes apesctos:

- Una introducción a los elementos esenciales de la Teoría de la Decisión, los cuales proporcionan, bajo el concepto de probabilidad subjetiva (grado de creencia), la base axiomática que consolida la lógica del enfoque Bayesiano.

- La presentación de los principios básicos de la inferencia Bayesiana, descando el concepto de intercambiabilidad, el Teorema de Repesentación y el empleo de la regla de Bayes bajo la expresión:
$$\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta).$$

- Una profundización sobre los métodos expuestos en la literatura para la elección de la distribución incial y, así mismo, sobre las técnicas ultilzadas actualmente para obtener la distribución posterior, enfatizando en los métodos MCMC a través de los algoritmos Metropolis-Hastings y Muestreador Gibbs (programados en diferentes softwares de plataforma libre) y, sobre las diferentes pruebas de diagnostico respecto a la convergencia de una cadena.

- El empleo del enfoque Bayesiano en el análisis de un problema de gran importancia en la actualidad referido a la predicción de dióxido de carbono en suelos de la zona de Veracuz, México, problema relacionado intimamente con el tema del cambio climático. Específicamente, se presenta una modelación con tres distribuciones iniciales distintas concluyendo, después de un analisis comparativo básico, que la relación entre las variables de interés, es descrita adecuadamente por el modelo que emplea la distribucion incial conjugada.

Apéndice A

Enfoque Bayesiano de los procedimientos de la estadística

El presente apéndice, expone brevemente, las contrapartes Bayesianas de los procedimientos inferenciales, esto es, los procedimientos de *estimación* y *pruebas de hipótesis*.

La inferencia Bayesiana[1] puede ser descrita técnicamente como un problema de decisión donde el espacio de acciones es la clase de las distribuciones de probabilidad final de la cantidad de interés. Puesto que la metodología frecuencial siempre ignora la información inicial, las distribuciones finales proporcionan la única forma de comparar los resultados estadísticos clásicos con los resultados Bayesianos: generalmente, los resultados clásicos pueden ser reproducidos a partir de las distribuiciones finales, por lo que, pragmática, *aunque no conceptualmente*, el paradigma Bayesiano constituye una *extensión* de la metodología clásica [7].

A.1. Naturaleza de la Inferencia Bayesiana

Generalmente el análisis estadístico para determinados datos observacionales D, comienza con alguna evaluación *descriptiva* informal, la cual es empleada para proponer, tentativamente, un *modelo de probabilidad*:

$$\{p(D \mid \omega), \omega \in \Omega\},$$

que para algún valor (desconocido) de ω, se asume representa el mecanísmo probabilístico que ha generado los datos D. Los argumentos presentados en los capítulos 1 y 2, establecen la necesidad lógica para considerar una distribución inicial sobre el espacio paramétrico Ω- $p(\omega)$, la cual describe el conocimiento inicial sobre el valor de ω antes de que los datos sean observados. De la teoría de la probabilidad se sigue que, si el modelo es el correcto, toda la información disponible sobre el valor ω, una vez observados los datos D, estará contenida en la distribución condicional:

$$p(\omega \mid D) = \frac{p(D \mid \omega)p(\omega)}{\displaystyle\int_{\Omega} p(D \mid \omega)p(\omega)d\omega}, \tag{A.1.1}$$

que también puede escribirse como:

$$p(\omega \mid D) = cp(D \mid \omega)p(\omega) \quad \text{con } c = [\int_{\Omega} p(D \mid \omega)p(\omega)d\omega]^{-1}. \tag{A.1.2}$$

[1] Ver Sección 1.3.

La afirmación (A.1.1) (equivalentemente (A.1.2)), es citada como el *Teorema de Bayes*, donde $p(\omega)$ es conocida como la distribución *inicial* y $p(\omega \mid D)$ como la distribución *final* [2] de ω, ambas expresan el conocimiento que se tiene sobre ω antes y después de la observación de los datos. La cantidad **c**, simplemente, es una constante de *proporcionalidad*, necesaria para asegurar que $p(\omega \mid D)$ sea una distribución de probabilidad; este hecho, permite prescindir de ésta constante para facilitar el cálculo de la posterior, así, el teorema de Bayes puede ser, simplemente, expresado como:

$$p(\omega \mid D) \propto {}^{3}p(D \mid \omega)p(\omega). \tag{A.1.3}$$

Para identificar la forma de la posterior, una técnica es identificar el *Kernel* de la densidad de probabilidad correspondiente ($k(\omega)$) y una función que no involucre a ω ($c(D)$), de tal forma que:

$$p(\omega \mid D) = c(D)k(\omega).$$

Cuando $p(D \mid \omega)$ es considerada como una función no de D pero si de ω, $p(D \mid \omega) = l(\omega \mid D)$ es conocida como la *función de verosimilitud*, así, el teorema de Bayes afirma que: *La posterior es proporcional a la verosimilitud por la inicial* [15]:

$$\text{Posterior} \propto \text{Verosimilitud} \times \text{Inicial}.$$

La función de verosimilitud desempeña un papel importante en la fórmula de Bayes, ya que a través de ella, los datos modifican el conocimiento inicial sobre ω, por lo tanto, es considerada como la representación de la información sobre ω, a partir de los datos. Además, de la ecuación (A.1.3), se sigue que si la misma inicial $p(\omega)$ es empleada para dos diferentes conjuntos de datos D_1 y D_2 con posiblemente, diferentes modelos de probabilidad $p_1(D_1 \mid \omega)$ y $p_2(D_2 \mid \omega)$ pero con función de verosimilitud *proporcional*, produciran idénticas posteriores para ω. Esta consecuencia inmediata del teorema de Bayes ha sido considerada como el *principio de verosimilitud* y, es considerada como un requerimiento obvio para la inferencia estadística, sin embargo, el principo de verosimilitud sólo es aplicable para inferencias sobre el vector paramétrico ω una vez que los datos han sido obtenidos.

Los términos inicial y posterior son relativos para un conjunto de datos particular. De la coherencia inducida por la teoría de la probabilidad, se sigue que si los datos $D = \{x_1, ..., x_n\}$ son presentados secuencialmente, el resultado final será el mismo si se procesan global o secuencialmente los datos, es decir:

$$p(\omega \mid x_1, ..., x_{i+1}) \propto p(x_{i+1} \mid \omega)p(\omega \mid x_1, ..., x_i), \quad \text{para } i = \overline{1, n-1}.$$

Por lo que, la *posterior* en un determinado momento se convierte en la *inicial*. Cabe señalar que en muchas situaciones, la posterior es más 'nítida' que la inicial, es decir, en muchos casos la densidad $p(\omega \mid x_1, ..., x_{i+1})$ estará más próxima al verdadero valor de ω que $p(\omega \mid x_1, ..., x_i)$. Sin embargo, esto no siempre es el caso: ocasionalmente, una observación aberrante incrementará, más que decrementar, la incertidumbre sobre el valor de ω.

Para un modelo de probabilidad dado, se puede hallar alguna función particular de los datos, $t(D)$, que sea *suficiente* en el sentido estadístico. Dado el modelo, $t(D)$ contiene toda la información sobre ω la cual está disponible en D, formalmente, $t = t(D)$ es suficiente si y sólo si existen funciones no-negativas f y g tal que la función de verosimilitud puede ser factorizada en la forma $l(D \mid \omega) = f(\omega, t)g(D)$. Note que una estadística suficiente siempre existe, $t(D) = D$ es obviamente suficiente; sin embargo, una estadística suficiente mucho más simple, con una dimensionalidad fija que es independiente del tamaño de muestra, frecuentemente existe. En efecto, es conocido, que será el caso si el modelo de probabilidad pertenece a la *familia exponencial generalizada*[4]. Es fácil establecer que si t es suficiente, la posterior de ω sólo depende de los datos D a través de $t(D)$, y

[2] Es común también referirse a estas distribuciones como la distribución *a priori* y *a posteriori*, respectivamente.
[3] $\propto$- denotará *porporcional a*.
[4] Familia que incluye algunos de los modelos probabilísticos más frecuentemente usados.

puede ser directamente calculada en términos de $p(t \mid \omega)$, por lo tanto, la ecuación (A.1.3) puede ser re-escrita como:

$$p(\omega \mid D) = p(\omega \mid t) \propto p(t \mid \omega)p(\omega).$$

También, el teorema de Bayes, contínua siendo técnicamente válido si $p(\omega)$ es reemplazada por una *función inicial impropia* $\pi(\omega)$[5]. Si la constante de proporcionalidad existe, la distribución inicial impropia[6] lleva a una posterior *propia* bien definida, así:

$$\pi(\omega \mid t) \propto p(t \mid \omega)\pi(\omega)^{7}.$$

A.1.1. Estimación

Desde el punto de vista Bayesiano, el resultado final de un problema de inferencia sobre alguna cantidad desconocida no es posible sin la correspondiente distribución posterior. Así, dado un conjunto de datos D, todo lo que podemos decir sobre alguna función del parámetro ω, que rige el modelo, estará contenida en la distribución posterior $p(\omega \mid D)$. Como se ha mencionado anteriormente, la inferencia Bayesiana puede ser descrita técnicamente como un problema de decisión donde el espacio de acciones es la clase de éstas distribuciones de probabilidad posterior (de la cantidad de interés) que son compatibles con las suposiciones aceptadas.

Sin embargo, para hacer más fácil la asimilación, es frecuentemente conveniente *resumir* la información contenida en la distribución posterior por: (i) proponiendo valores para la cantidad de interés que, a la luz de los datos, son verosímiles a estar más *cerca* al verdadero valor de ésta y, (ii) midiendo la compatibilidad de los resultados con los valores hipotéticos de la cantidad de interés, la cual pudo haber sido sugerida por el contexto de la investigación.

Estimación Puntual

Sea D un conjunto de datos, asumiendo que han sido generados por un modelo de probabilidad $\{p(D \mid \omega), \omega \in \Omega\}$, y sea $\theta = \theta(\omega) \in \Omega$ la cantidad de interés. Una *estimación puntual* de θ es una función de los datos $\tilde{\theta} = \theta(D)$, la cual será considerada como una apropiada aproximación al verdadero, pero desconocido valor de θ. Formalmente, la elección de una estimación para θ es un *problema de decisión*, donde el espacio de decisiones es la clase Θ de posibles valores de θ [12]. Desde la perspectiva de la teoría de la decisión, la elección de una estimación puntual $\tilde{\theta}$, de alguna cantidad θ, es una decisión que actúa como si $\tilde{\theta}$ fuese θ, la cual no afirma algo sobre el valor de θ (aunque se desea afirmar algo simple, puede ser bien la razón para obtener una estimación). Tal como se estipula por la fundamentación de la teoria de la decisión, para resolver este problema de decisión es necesario especificar una *función de pérdida* $L(\tilde{\theta}, \theta)$, la cual mide la consecuencia de actuar como si el valor verdadero de la cantidad de interés fuese $\tilde{\theta}$ cuando es realmente θ. La pérdida posterior esperada, si $\tilde{\theta}$ fuese empleada es:

$$\overline{L}[\tilde{\theta} \mid D] = \int_{\Theta} L(\tilde{\theta}, \theta)p(\theta \mid D)d\theta$$

y el correspondiente *estimador de Bayes* es la función de los datos, $\theta^* = \theta^*(D)$, que minimiza esta expresión.

De manera más formal [14], un estimador Bayesiano puntual queda definido por:

[5] $\pi(\omega)$ es una función incial impropia, si es una función positiva tal que $\int_{\Omega} \pi(\omega)$ es no finita.

[6] $\pi(\omega)$ también puede ser una función inicial 'no-informativa' (generalmente impropia)-Ver Secc. 2.3.1.

[7] $p(\omega \mid t) = \pi(\omega \mid t)$.

Definición A.1.1. Un *estimador Bayesiano* de θ con respecto a la función de pérdida $L(\cdot, \theta)$ y la distribución $p(\theta)$ es un $\tilde{\theta} \in \Theta$ que minimiza:

$$\int_{\Omega} L(\tilde{\theta}, \theta) p(\theta) d\theta.$$

Algunas formas de los estimadores Bayesianos, dependiendo la función de pérdida elejida, son:

Proposición A.1.1. *1. Si $\Theta = \Re^k$ y $L(\tilde{\theta}, \theta) = (\tilde{\theta} - \theta)^t H(\tilde{\theta} - \theta)$ con H simétrica definida positiva entonces, el estimador de Bayes satisface:*

$$H\tilde{\theta} = HE(\theta).$$

Si H^{-1} existe entonces, $\tilde{\theta} = E(\theta)$, es decir, el estimador de Bayesiano, con respecto a la forma cuadrática de pérdida, es la media de $p(\theta)$, asumiendo que la media existe.

2. Si $\Theta = \Re$ y $L(\tilde{\theta} - \theta) = c_1(\tilde{\theta} - \theta) I_{\theta \leq \tilde{\theta}}(\theta) + c_2(\tilde{\theta} - \theta) I_{\theta > \tilde{\theta}}$ entonces, el estimador Bayesiano, con respecto a pérdida lineal, es el cuantil tal que:

$$P(\theta \leq \tilde{\theta}) = \frac{c_2}{c_1 + c_2}.$$

Si $c_1 = c_2$ el estimador de Bayesiano, con respecto al valor absoluto de la pérdida, es la mediana de $p(\theta)$.

3. Si $\Theta \subseteq \Re^k$ y $L(\tilde{\theta} - \theta) = 1 - I_{(B(\theta))}(\theta)$, donde $B(\theta) \subseteq \Theta$ es un una bola con centro en θ y radio ϵ entonces, el estimador de Bayesiano es el que maximiza:

$$\int_{Be(\theta)} p(\theta) d\theta.$$

Cuando $\epsilon \to 0$, la función que maximiza tiende a $p(\theta)$ así, el estimador Bayesiano, con respecto a la pérdida cero-uno, es la moda de $p(\theta)$, asumiendo que la moda existe.

Estimación por Intervalo

Para describir el contenido inferencial de la distribución posterior de la cantidad de interés $p(\theta \mid D)$ es, frecuentemente, conveniente dar regiones $R \subset \Theta$ de probabilidad dadas en virtud de $p(\theta \mid D)$ [14, 29]. Por ejemplo, la identificación de regiones que contienen el 50 %, 90 %, 95 % ó 99 % de la probabilidad. Una región $R \subset \Theta$ tal que $\int_R p(\theta \mid D) = q$ (de modo que, dados los datos D, el valor verdadero de θ pertenezca a R con probabilidad q), se dira que es una *región q-creible* posterior de θ. Note que esta proporciona inmediatamente una afirmación directa intuitiva sobre la cantidad desconocida de interés, θ, en términos de probabilidad, en marcado contraste con las afirmaciones circunlocutorias previstas por los intervalos de confianza frecuentistas. Claramente, para alguna q dada hay generalmente una infinidad de regiones de credibilidad[8] pero, algunas veces, las regiones de credibilidad son seleccionandas para tener un tamaño mínimo, dando por resultado un región con densidad de probabilidad alta (HPD), donde todos los puntos en la región tienen una densidad de probabilidad mayor que todos los puntos fuera. Sin embargo, las regiones HPD, no son invariantes bajo reparametrizaciones: la imagen $\phi(R)$ de una región R de HPD, será una región de credibilidad para ϕ, pero no generalmente será HPD. Los cuantiles a posteriori son frecuentemente usados para derivar regiones de credibilidad. Así, si $\theta_q = \theta_q(D)$ es el cuantil posterior $100q$ % de

[8] Una región de credibilidad es invariante bajo reparametrizaciones por lo que, si para alguna región q-créible R de θ, $\phi(R)$ será una región q-créible de $\phi = \phi(\theta)$.

θ entonces, $R = \{\theta; \theta \leq \theta_q\}$ es una región q-créible unilateral, generalmente única. Más aún, la regiónes q-ceíbles de *probabilidad centrada* de la forma $R = \{\theta; \theta_{\frac{1-q}{2}} \leq \theta \leq \theta_{\frac{1+q}{2}}\}$ son fácilmente calculables y son frecuentemente más empleadas que las regiones HPD.

Análogamente a la estimación puntual, más formalmente una región de credibilidad Bayesiana queda definida por [14]:

Definición A.1.2. Una región $R \subseteq \Theta$ tal que:

$$\int_R p(\theta) d\theta = 1 - \alpha.$$

Se dira que es una $100(1-\alpha)\,\%$ región de credibilidad para θ, con respecto a $p(\theta)$.

Claramente, para todo α no existe una única región de credibilidad, por lo tanto, el problema de elegir entre subjuntos $R \subseteq \Theta$ tal que $\int_R p(\theta) d\theta = 1 - \alpha$, para un Θ, $p(\theta)$ y un α (fijo) dados; puede ser visto como un problema de decisión siempre que se especifique una función de pérdida, $L(R, \theta)$, que refleje las posibles consecuencias de elegir la $100(1-\alpha)\,\%$ región R.

Proposición A.1.2. *Sea $p(\theta)$ una distribución de probabilidad continua para toda $\theta \in \Theta$ excepto en una cantidad finita de puntos. Sea $0 < \alpha < 1$ y $\mathcal{A} = \{R : P(\theta \in R) = 1 - \alpha\} \neq \emptyset$ y $L(R, \theta) = k \parallel R \parallel - I_{R(\theta)}$ con $R \in \mathcal{A}$, $\theta \in \Theta$ y $k > 0$ entonces, R es optima si y sólo si tiene la siguiente propiedad:*

$$\forall \theta_1 \in R, \theta_2 \notin R : p(\theta_1) \geq p(\theta_2),$$

excepto posiblemetne en un subconjunto de Θ de probabilidad cero.

Intuitivamente la proposición A.1.2 describe que la región de credibilidad a elegir, para un α fijo y para una función de pérdida seleccionada, es aquella con $\parallel R \parallel$ mínima.

Así, de la proposición A.1.2, se tiene más formalmente lo que será una región con densidad de probabilidad alta:

Definición A.1.3. Una región $R \subseteq \Theta$ será llamada una $100(1-\alpha)\,\%$ región con densidad de probabilidad alta para θ con respecto a $p(\theta)$ si:

1. $P(\theta \in R) = 1 - \alpha$.
2. $\forall \theta_1 \in R, \theta_2 \notin R : p(\theta_1) \geq p(\theta_2)$ excepto posiblemente para un conjunto de Ω con probabilidad cero.

A.1.2. Pruebas de Hipótesis

La distribución posterior $p(\theta \mid D)$, de la cantidad de interés θ, transmite de inmediato información intuitiva sobre los valores de θ que, dadas las suposiciones del modelo, pueden ser tomadas para ser *compatibles* con los datos observacionales D, es decir, aquellos con una alta densidad de probabilidad. Algunas veces, una resticción $\theta \in \Theta_0 \subset \Theta$, de los posibles valores de la cantidad de interés (donde Θ_0 puede posiblemente consistir de un valor singular θ_0) es sugerida como una consideración especial, ya sea porque restringuiendo Θ a Θ_0 simplificaría en gran medida el modelo, o porque existen además otros argumentos específicos en el contexto que sugieren que $\theta \in \Theta_0$ [7, 12, 31]. Intuitivamente, la *hipótesis* $H_0 \equiv \{\theta \in \Theta_0\}$ debe ser considerada *compatible* con los datos D si existen elementos en Θ_0 con una, relativamente, alta densidad posterior. Sin embargo, una conclusión más precisa es frecuentemente requerida y, una vez más, esto es posible hacerlo adoptando un enfoque orientado hacia la teoría de la decisión. Formalmente, una prueba de hipótesis $H_0 \equiv \{\theta \in \Theta_0\}$ es un *problema de decisión* donde el espacio de acciones sólo tiene dos elementos, aceptar (a_0) o rechazar (a_1) la restricción propuesta. Para resolver este problema

de decisión, es necesario especificar una apropiada función de pérdida $L(a_i, \theta)$, la cual mida las consecuencias de aceptar o rechazar H_0 como una función del verdadero valor de θ. Note que esto requiere la especificación de una *alternativa* a_1 ha aceptar H_0; ésto sólo será esperado, para una acción tomada no porque sea buena, sino, porque es mejor a todo lo demás.

Dados los datos D, la acción optima será rechazar H_0 si y sólo si la pérdida posterior esperada de aceptación:

$$\int_{\Theta} L(a_0, \theta) p(\theta \mid D) d\theta,$$

es más grande que la pérdida posterior esperada de rechazar:

$$\int_{\Theta} L(a_1, \theta) p(\theta \mid D) d\theta,$$

esto es, sí (y sólo si):

$$\int_{\Theta} [L(a_0, \theta) - L(a_1, \theta)] p(\theta \mid D) d\theta = \int_{\Theta} \Delta L(\theta) p(\theta \mid D) d\theta > 0.$$

Por tanto, sólo la diferencia de pérdida $\Delta L(\theta) = L(a_0, \theta) - L(a_1, \theta)$, mide la *ventaja* de rechazar H_0 como una función de θ, por lo que debe ser especificada. Así, la hipótesis H_0 debería ser rechazada si la ventaja esperada de rechazar H_0 es positiva.

Un elemento crucial en la especificación de la función de pérdida es la descripción de lo que realmente se entenderá por rechazar H_0. Por supuesto que a_0 significa actuar como si H_0 fuese cierta, es decir, como si $\theta \in \Theta_0$, pero hay al menos dos significados para la acción alternativa a_1. Esto puede significar:

- La *negación* de H_0, esto es, como si $\theta \notin \Theta_0$.

o, alternativamente:

- Rechazar la simplificación implicada por H_0 y mantener el modelo sin restricciones, $\theta \in \Theta$, lo cual es cierto por suposición.

Ambas alternativas han sido analizadas en la literatura, aunque se puede argumentar que los problemas de análisis de datos, donde los procedimientos de pruebas de hipótesis son generalmente usados, son mejor descritos por la segunda alternativa. Más aún, si un modelo establecido es identificado por $H_0 \equiv \{\theta \in \Theta_0\}$ entonces, es frecuentemente incrustado dentro de un modelo más general, $\{\theta \in \Theta, \Theta_0 \subset \Theta\}$, construído así para incluir posiblemente salidas prometedoras de H_0, y esto es requerido para verificar si realmente los datos D siguen siendo compatibles con $\theta \in \Theta_0$, o si la extensión de $\theta \in \Theta$ es realmente requerída.

Apéndice B

Predicción de dióxido de carbono en Veracruz

Los estudios sobre el dióxido de carbono (uno de los gases causantes del efecto invernadero) son de suma importancia debido a su relación con el fenómeno del calentamiento global. Estos gases son continuamente emitidos y removidos en la atmósfera por procesos naturales sobre la Tierra, pero las actividades antropogénicas causan cantidades adicionales de los mismos, incrementando sus concentraciones en la atmósfera, lo que tiende a sobrecalentarla.

En la actualidad, existe gran interés científico en modelar la relación que guardan las emisiones del dióxido de carbono (CO_2) con las propiedades del suelo y, con el propósito de estudiar estas asociaciones, el presente apéndice muestra el análisis de datos efecuado a través del modelo de regresión lineal simple, desde el enfoque clásico.

B.1. Enfoque Clásico

La evaluación de las emisiones de CO_2 y las propiedades del suelo, en determinados sitios de muestreo del Parque Ecológico Jaguaroundi en Coatzacoalcos, Veracruz, México, se efectúo utilizando el método de cámara estática. Ubicando homogeneidad en el terreno y considerando un mismo plano por altitud y cubierta vegetal, se colocaron 3 cámaras en los sitios seleccionados. Las emisiones se analizaron con un Detector de Ionización de Flama y las propiedades físicas y químicas de las muestras de suelo, fueron estudiadas por el Grupo de Edafología del Instituto de Geología de la UNAM [36].

Los 24 datos muestrales a analizar estan dados en la tabla B.1.

Datos					
CO_2	321.890	283.620	425.380	-228.000	531.500
	273.060	10.526	9.487	-23.339	36.291
	-8.039	25.954	304.231	100.039	130.866
	59.759	126.388	99.571	44.376	101.781
	53.160	148.698	-15.208	156.298	
Nitrógeno	0.675	0.480	1.130	0.170	0.265
	0.230	0.115	0.240	0.285	0.245
	0.190	0.255	0.185	0.220	0.160
	0.280	0.260	0.280	0.105	0.150
	0.165	0.135	0.105	0.135	

Tabla B.1: Mediciones de emisiones de CO_2 y Nitrógeno

El análisis de datos a través del modelo de regresión lineal simple, desde el enfoque clásico,

hace necesario tener la suposición de normalidad en los errores y homogeneidad en las varianzas.

Para el caso de la estimación clásica, se ha empleado el sotfware SPSS [40], por lo que, el análisis de regresión lineal simple de la variable CO_2 sobre la variable Nitrógeno efectuado a través de dicho sotfware, arrojo los siguientes resultados (Fig. B.1).

Tabla 1a: Coeficientes del modelo de regresión lineal para las emisiones de CO2

Modelo		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Constante	15.121	45.835		.330	.745
	Nitrógeno	403.312	132.568	.544	3.042	.006

a. Dependent Variable: CO2

Tabla 1b: Resumen del modelo

Modelo	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.544[a]	.296	.264	140.93291

a. Predictors: Constante, Nitrógeno
b. Dependent Variable: CO2

Tabla 1c: ANOVA

Modelo		Suma de cuadrados	df	Media cuadrática	F	Sig.
1	Regresión	183836.1	1	183836.110	9.256	.006[a]
	Residual	436965.8	22	19862.084		
	Total	620802.0	23			

a. Predictors: Constante, Nitrógeno
b. Dependent Variable: CO2

Figura B.1: Resultados

Como puede observarse, en las Tablas 1a, 1b y 1c, la asociación lineal entre la variable CO_2 y Nitrógeno es estadísticamente significativa[1] y se puede concluir que:

A mayor Nitrógeno, mayor emisión de CO_2.

En la regresión lineal se supone que los verdaderos errores son independientes con distribución $N(0, \sigma^2)$. Los residuos son la estimación a los verdaderos errores, y la estimación de σ^2, s^2, es el cuadrado medio residual, donde s es el error típico de la estimación. Para este caso en partícular, el error típico de la estimación es de 140.93 (Tabla 1b-Resumen del modelo) y el valor de R^2 es bastante bajo, 29.6 %, sin embargo, en la Tabla 1c, se muestra que el estadístico F es significativo, indicando que el predictor nitrógeno puede explicar a la variable respuesta (CO_2).

Respecto a la normalidad, la distribución de la variable formada por los residuos debe ser normal: los residuos observados y los esperados bajo la hipótesis de distribución Normal deben ser parecidos. Para comprobar esto, una alternativa es emplear el gráfico de probabilidad normal, que permite comparar, gráficamente, la función de distribución observada en la muestra con la función de distribución de una Normal $(0, 1)$. El gráfico (Fig. B.2) de probabilidad normal para los residuos tipificados representa la función de distribución esperada bajo la hipótesis de distribución Normal $(0, 1)$, eje vertical, frente a la distribución acumulativa de los valores observados, eje horizontal. Como puede observarse, la distribución de los residuos es aproximadamente normal, debido a que los valores de los residuos son aproximadamente iguales a la recta que pasa por el origen con pendiente igual a 1. Luego, se puede aceptar que los residuos proceden de una distribución Normal.

Con respecto a la independencia, el valor observado no debe estar influenciado en ningún sentido por los demás valores observados: los residuos no deben presentar ningún patrón sistemático

[1] $p = 0{,}006$.

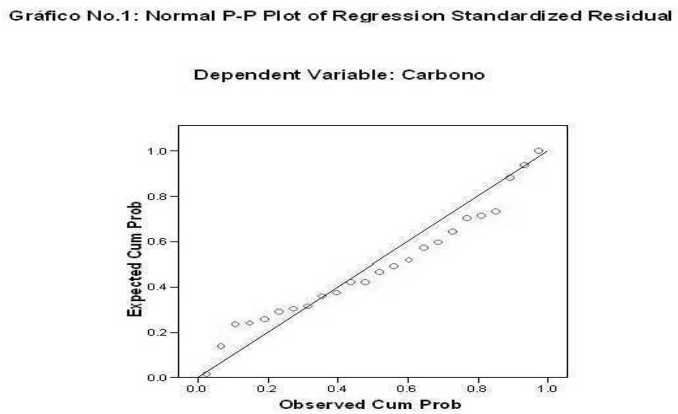


Figura B.2: Análisis de Residuos

respecto a la secuencia de observaciones. El estadístico de Durbin-Watson, D, mide el grado de autorrelación entre el residuo correspondiente a cada observación y la anterior. Si su valor es próximo a 2, los residuos estarán incorrelados; si se aproxima a 4, estarán negativamente autocorrelados, y si se aproxima a 0, estarán positivamente autocorrelados. En nuestro caso, el valor del estadístico de Durbin-Watson es igual a 2,18, que es próximo a 2, lo que confirma la incorrelación de los residuos.

Por último, las varianzas de las distribuciones de la variable dependiente ligadas a los distintos valores de la variable independiente deben ser iguales: los residuos no deben presentar ningún patrón sistemático respecto de las predicciones o respecto a la variable independiente. Para analizar la homogeneidad de varianzas se empleó el gráfico de residuos tipificados (Figura B.3), frente a las estimaciones tipificadas. Si la varianza de los resiudos fuera constante, la nube de puntos estará concentrada en una banda, centrada en el cero y paralela al eje de las abscisas. Para este caso, se puede observar que la nube de puntos se encuentra, en efecto, centrada en el cero y paralela al eje de las absisas, por lo que gráficamente podemos afirmar que la suposición de homogeneidad de varianzas se satisface.

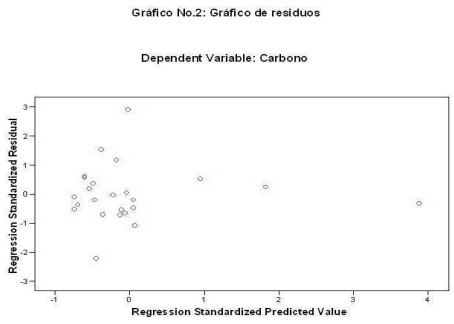


Figura B.3: Residuos

Así, el modelo de regresión lineal simple, analítica[2] y gráficamente, desde el enfoque clásico está dado conjuntamente en la Figura B.4:

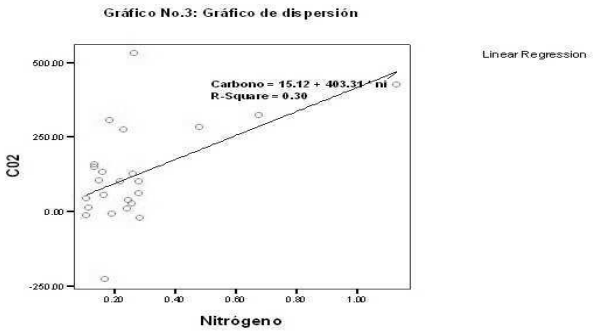


Figura B.4: Modelo de Regresión

Zelig

El paquete informático Zelig, permite realizar estimaciones máximo verosímiles desde el enfoque clásico. La especificación[3] *model = 'normal'*, analiza el modelo de regresión lineal normal.

- Codificación del modelo:
  - $CO_2 \sim N(\mu_i, \tau^{-1})$.
  - $\mu_i = \beta Ni_i$.
- Inferencias:
  - El resumen de los coeficientes de regresión, es decir, las estimaciones máximo verosímiles de los parámetros del modelo, se encuentra dada en el Figura B.5 :

```
> Posterior<- zelig(CO2 ~ NI, model="normal", data=datos)
> summary(Posterior)

Call:
zelig(formula = CO2 ~ NI, model = "normal", data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-311.68   -73.92   -20.78    57.14   409.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.12      45.84   0.330  0.74460
NI            403.31     132.57   3.042  0.00598 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 19862.08)

    Null deviance: 620802  on 23  degrees of freedom
Residual deviance: 436966  on 22  degrees of freedom
AIC: 309.54

Number of Fisher Scoring iterations: 2
```

Figura B.5: Resumen de los Coeficientes de Regresión

[2]La estimación para los parámetros del modelo están dados en la Fig.B.1: Tabla 1.a.
[3]Ver Sección 2.4.1.

Bibliografía

[1] Alonso R. A. (2008). *El resurgir de Thomas Bayes.* Anuario Jurídico y Económico Escurialense **XLI**, 327-360.

[2] Berger, J. O. (1980). *Statistical Decision Theory and Bayesian Analyis.* Springer, Second Edition.

[3] Berger, J. O. and Bernardo, J. M. (1992). *On the development of reference priors.* Bayesian Statistics 4 (J. M. Bernardo, J. O. Berger, D. V. Lindlye and A. F. M. Smith, eds). Oxford: Oxford University Press, 35-60 (with discussion).

[4] Berger, J. O., Bernardo, J. M. and Sun, D. (2009).*The formal definition of Reference priors.* Annals of Statistics **37**, 905-938.

[5] Bernardo, J. M. (1979). *Expected information as expected utility.* The Annals of Statistics **7**, 686-690.

[6] Bernardo, J. M. (1981). *Bioestadística: Una perspectiva Bayesiana.* Vicens-Vives.

[7] Bernardo, J. M.(1985). *Análisis Bayesiano de los contrastes de hipótesis paramétricos.* Trabajos de Estadística **36**, 45-54.

[8] Bernardo, J. M. (1989). *Análisis de Datos y Metodos Bayesianos.* Historia de la Ciencia Estadística (S. Ríos, ed). Madrid: Real Académia de Ciencias, 87-105. Reprinted in *Pro Mathematica* **2** (1989), 27-55.

[9] Bernardo, J. M. (1997). *Bruno de Finetti en la Estadística Contemporánea.* Historia de la Matemática en el Siglo XX (S. Ríos, ed). Madrid: Real Academia de Ciencias, 63-80.

[10] Bernardo, J. M. (1997). *Noninformative Priors Do Not Exist: A discussion.* J. Statistics. Planning and Inference **65**, 159-189 (with discussion).

[11] Bernardo, J. M. (1996). *The Concept of Exchangeability and its Applications.* Far East J. Mathematical Sciences **4**, 111-121.

[12] Bernardo, J. M. (2003). *Bayesian Statistics.* Encyclopedia of Life Support Systems (EOLSS), Probability and Statistics (R. Viertl, ed). Oxford, UK: UNESCO (www.eolss.net).

[13] Bernardo, J. M. (2005). *Reference Analysis.* Handbook of Statistics **25** (D. K. Dey and C. R. Rao eds.). Amsterdam: Elsevier, 17-90.

[14] Bernardo, J. M. and Smit, A. F. M. (1994). *Bayesian Theory.* Wiley.

[15] Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis.* Wiley.

[16] Congdom, P. (2003). *Applied Bayesian Modelling.* Wiley.

[17] Cruz, M. E. (2006). *WinBUGS: un software para Inferencia Bayesiana.* Tesis de licenciatura, UNAM, México.

[18] Florens, J. P., Mouchart, M. and Rolin, J. M. (1990). *Elements of Bayesian Statistics.* CRC Press.

[19] Gelman, A. (2002). *Prior distribution.* Encyclopedia of Environmetrics. **3**, 1634-1637.

[20] Gelman, A. (2009). *Bayes, Jeffreys, piror distributions, and the philosophy of statistics.* Statistical Science **24**, 176-178.

[21] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis.* Champman & Hall, Second Edition.

[22] Imai K., King G. and Lau O. (2008). *Zelig: Everyone's Statistical Software.* R package version 3.4-8, URL http://gking.harvard.edu/zelig.

[23] Lee, P. M. (1997). *Bayesian Statistics an Introduction.* Hodder Arnold.

[24] Lindley, D. V. (1983). *Theory and Pactice of Bayesian Statistics.* Journal of the Royal Statistical Society. Series D (The Statistician) **32**, No. 1/2, 1-11.

[25] Lindley, D. V. (1987). *Bayesian Statistics: A Review.* Society for Industrial and Mathematics.

[26] Lindley, D. V. (2000). *The Philosophy of Statistics.* Journal of the Royal Statistical Society. Series D (The Statistician), **49**, No. 3, 293-337.

[27] Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientistis.* Springer.

[28] Martin A.D., Quinn K.M. (2007). *MCMCpack: Markov chain Monte Carlo (MCMC) Package.* R package version 1.0-7, URL http://mcmcpack.wustl.edu.

[29] Migon, H. and Gamerman, D. (1999). *Statistical Inference an Integrated Approach.* Hodder Arnold.

[30] Montgomery, D. C., Peck, E. A. and Vining, G. G. (2002). *Introducción al Análisis de Regresión Lineal.* Compañia Editorial Continental.

[31] Moreno, E. and Girón, F. J. (2006). *On the frequentist and Bayesian approaches to hypothesis testing.* Statistics & Operations Research Transactions, 3-28.

[32] Press, S. J. (1989). *Bayesian Statistics: Principles, Models, and Applications.* Wiley.

[33] Press, S. J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications.* Wiley, Second Edition.

[34] Robert, C. P. (2007). *The Bayesian Choice: From Descison-Theoretic Foundations to Computational Implementation.* Springer.

[35] Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods.* Springer.

[36] Saldaña, M. J. A. y Ruiz, S. L. G. (2007). *Emisiones de gases de efecto invernadero en suelos perturbados con diferente cobertura vegetal en Coatzacoalcos, Veracruz, México.* (Artículo por publicar).

[37] Sánchez M., Gómez E. and Gómez, V. M. A. (1990). *Three Methods for Constructing Reference Prior Distributions.* Rev. Mat. Univ. Complut. Madrid **3**, 153-162.

[38] Smith, B. J. (2004). *Bayesian output analysis program (BOA) version 1.1 user's manual.* Technical report, University of Iowa. http://www.public-health.uiowa.edu/boa.

[39] Sparapani, R. A., (Prakash) Laud P. W. (2008). *A Recent History of Bayesian Statistical Software.* The American Statistical Association, Milwaukee Chapter of the ASA (MILWASA): Introduction Workshop on Bayesian Medeling an Inference. March, Milwaukee, Wisconsin.

[40] SPSS 15.O for Windows Evaluation Versión.

[41] Suppes, P. (2007). *Where do Bayesian priors come from?.* Synthese **156**, 441-471.

[42] WinBUGS with Doodle Bugs, Versión 1.4.

[43] Zhu, M. (2004). *The Counter-intuitive Non-informative Prior for the Bernoulli Family.* Journal of Statistics Education **12**, 1-10.