Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Regresión Lineal por Medio del Análisis Bayesiano

Tesis presentada al

Colegio de Matemáticas

como requisito parcial para obtener el grado de

Licenciado en Matemáticas

por

Octavio Paredes Pérez

directores de tesis

Dra. Hortensia Reyes Cervantes
Dr. Francisco Tajonar Sanabria

Puebla Pue.
4 de Septiembre del 2013

Dedicado a mi familia

"La caida de los imperios que aspiran al dominio universal, puede predecirse con probabilidad muy alta por cualquiera versado en el cálculo del azar."

Pierre Simon de Laplace.

Gracias

Dra. Hortensia Reyes Cervantes,
Dr. Francisco Tajonar Sanabria.

" Por guiarme y enseñarme que el camino más dificíl de la vida, no es el de aprender, si no el de llevar a cabo lo aprendido."

Índice general

Capítulo 1

Introducción

El origen de está tesis, se dio por curiosidad de saber lo que es la estadística bayesiana, porque en la licenciatura de matemáticas de la BUAP, no se da está materia, así en un principio se pretendia que fuése un trabajo bibliográfico, dado la complejidad de la materia, pero con el tiempo se pensó en dar una sencilla aplicación de lo que es la estadística bayesiana, en particular, la regresión lineal bayesiana y compararla con la regresión lineal clásica.

En la segunda mitad del siglo $XVIII$ fue publicado el teorema de Bayes (1764), así llamado por el nombre del monje que lo desarrolló, en respuesta a los postulados de la inferencia Gausiana. El estudio clásico de las distribuciones de probabilidad o estadística Gausiana supone funciones de densidad simétricas y bien definidas, así como la ausencia de cualquier conocimiento previo por parte del investigador. Bayes, en la justificación de su teoría argumenta que los datos no necesariamente provenían de tales funciones de densidad, sino que probablemente eran generados por leyes probabilísticas sujetas a formas asimétricas y sesgadas. En tanto que el investigador conociera estas características, el procedimiento correcto de inferencia estadística debería incorporar, decía Bayes [12], esta información y de esta forma, contar con un marco probabilístico más apropiado para la inferencia estadística.

En la actualidad se reconoce la importancia que tiene la estadística bayesiana en distintas áreas del conocimiento, como son salud, biología, astronomía, economía etc. [6, 30, 15, 25] Dichas áreas de investigación científica utilizán la estadística bayesiana para la recopilación, presentación, análisis e interpretación de los datos, con el fin de ayudar a una mejor toma de decisión.

La estadística se ocupa fundamentalmente del análisis de datos que presentan variabilidad con el fin de comprender el mecanismo que los genera, o para ayudar en un proceso de toma de decisiones. En cualquier caso, existe una componente de incertidumbre involucrada, por lo que él estadístico se ocupa en reducir lo más posible esa componente, así como también en describírla en forma apropiada. Se supone que toda forma de incertidumbre debe decidirse por medio de modelos de probabilidad. En el caso de la estadística bayesiana se considera al parámetro sobre el cuál se debe inferir, como un evento incierto. Entonces, como nuestro conocimiento no es preciso y está sujeto a incertidumbre, se puede describir mediante una distribución a priori o inicial del parámetro θ, lo que hace que el parámetro tenga el carácter aleatorio.

La estadística bayesiana es una alternativa a la estadística clásica, en los problemas estadísticos típicos como son estimación de los parámetros, pruebas de hipótesis y predicción. La estadística bayesiana proporciona un completo paradigma de inferencia estadística y teoría de la decisión con incertidumbre, está basada en el teorema de Bayes, que es el que nos proporciona una forma

adecuada para incluir información inicial o a priori de los parámetros mediante una distribución $p(\theta)$, además de la información muestral $p(x|\theta)$ para así describir toda la información disponible del parámetro mediante la distribución a posteriori o final $p(\theta|x)$.

La teoría bayesiana plantea la solución a un problema estadístico desde el punto de vista subjetivo de la probabilidad, según el cuál, la probabilidad que un estadístico asigna a uno de los posibles resultados de un proceso, representa su propio juicio sobre la verosimilitud de que se tenga el resultado. Este juicio estará basado en opiniones e información acerca del proceso. La información que el estadístico tiene sobre la verosimilitud de los distintos eventos reelevantes al problema de decisión, debe ser cuantificada a través de una medida de probabilidad Θ.

A un problema específico se le puede asignar cualquier tipo de distribución a priori o inicial, ya que finalmente al actualizar la información a priori que se tenga acerca del parámetro, por medio del teorema de Bayes, se estará actualizando la distribución a posteriori del parámetro.

Cuando un investigador tiene un conocimiento previo a un problema, este puede cuantificarse en un modelo de probabilidad. Si los juicios de una persona sobre la verosimilitud relativa a ciertas combinaciones de resultados satisfacen ciertas condiciones de consistencia, se puede decir que sus probabilidades subjetivas se determinan de manera única.

La diferencia fundamental entre el modelo clásico y el bayesiano es que en este último los parámetros son considerados aleatorios, por lo que pueden ser cuantificados en términos probabilísticos. Por otro lado, es importante resaltar que la inferencia bayesiana se basa en probabilidades asociadas con diferentes valores del parámetro θ que podrían haber dado lugar a la muestra x que se observó. Por el contrario, la inferencia clásica se basa en probabilidades asociadas con las diferentes muestras x que se podrían observar para algún valor fijo, pero desconocido del parámetro θ. En relación con la obtención de estimaciones puntuales para los parámetros poblacionales, en el caso del modelo clásico, la estimación se interpreta como el valor de θ, que hace más probable haber obtenido la muestra observada, mientras en el modelo bayesiano, la estimación sera el valor de θ que, puesto que se ha observado x, sea mas verosímil o más creíble.

Las principales características que se le pueden atribuir a la teoría bayesiana son las siguientes:

1. Proporciona una manera satisfactoria de introducir explícitamente y de dar seguimiento a los supuestos sobre el conocimiento inicial o a posteriori.

2. La inferencia bayesiana no presenta problemas en la selección de estimadores y de intervalos de credibilidad.

3. El teorema de Bayes permite la actualización continua de la información sobre los parámetros de la distribución conforme se generan más observaciones.

4. A diferencia de la inferencia clásica, la bayesiana no requiere de la evaluación de las propiedades de los estimadores obtenidos en un muestreo sucesivo.

5. La probabilidad de un evento está dada por el grado de confianza o creencia que tiene un individuo sobre la ocurrencia del evento.

La principal objeción es que las conclusiones dependen de la selección específica de la aproximación inicial. Aunque para otros esto es lo interesante de la aproximación bayesiana. Sin embargo, se debe señalar que inclusive en inferencia clásica y además en investigaciones científicas, por lo general estos conocimientos iniciales son útilizados implícitamente. Cabe recalcar que los métodos bayesianos pueden ser aplicados a problemas que han sido inaccesibles a la teoría frecuentista tradicional.

La estructura del trabajo de tesis es la siguiente:

En el capítulo uno, se da una pequeña introducción de lo que es la estadística bayesiana, así como también un poco de como se hace inferencia respecto a la estadística clásica.

En el capítulo dos, se menciona los tipos de promedios o medidas de tendencia central, especialmente lo que es la moda y la mediana, que se ocupa mucho en estadística bayesiana, así como su diferencia con respecto a trabajar con la estadística clásica, también mencionaremos el uso de la probabilidad subjetiva, nos adentraremos un poco de lo que es la robustez.

En el capítulo tres, se da un panorama general de lo que es el análisis bayesiano, el teorema de Bayes, función de verosimilitud, las distintas formas de obtener la distribución a priori y la distribución a posteriori, así como la forma de hacer inferencia bayesiana.

En el capítulo cuatro, se menciona el uso de la forma de como estimar en estadística bayesiana, el uso de la teoría de decisión en este capitulo, la familia exponencial y su importancia, intervalos de credibilidad así como la diferencia entre intervalo de credibilidad y de confianza, por último mencionaremos a las pruebas de hipótesis.

El capítulo cinco, contiene el desarrollo del análisis de regresión lineal bayesiano, iniciando con los modelos lineal simple y múltiple, con una distribución a priori no informativa y enseguida el modelo de regresión con distribución a priori conjugada, se determinan las distribuciones a posteriori de los parámetros de regresión.

Por último, el capítulo seis, aborda todo lo concerniente a la aplicación de los modelos, se dan los resultados obtenidos haciendo una comparación entre los resultados clásicos y bayesianos, se presentan grafícas de las distribuciones a posteriori, historial, gráfico de cuántiles, autocorrelación.

Finalmente de presentar las conclusiones obtenidas a lo largo del trabajo.

Octavio Paredes Pérez
Facultad de Ciencias Físico Matemáticas
Benemérita Universidad Autónoma de Puebla
17 *de Julio del* 2013

Capítulo 2

Preliminares

En este capítulo se presentan algunas nociones preliminares que serán de utilidad para el estudio de la Estadística Bayesiana, así como la diferencia entre la Estadística Clásica y la Bayesiana, el uso de la probabilidad subjetiva en ésta área de trabajo así como también de la importancia del Teorema de Bayes y su aplicación, por lo que daremos una simple introducción a las probabilidades a priori y a posteriori, que es la parte esencial de la Estadística Bayesiana, (Veáse [18]).

2.1. Estadística

De manera muy general, puede decirse que la estadística es la disciplina que estudia los fenómenos inciertos (aleatorios), es decir, aquellos que no se pueden predecir con certeza. El estudio se lleva a cabo a partir del posible conocimiento previo sobre el fenómeno y de observaciones que se realizan sobre el mismo.

2.1.1. Promedios o medidas de tendencia central

Un promedio es un valor típico o representativo de un conjunto de datos. Como tales valores suelen situarse hacia el centro del conjunto de datos ordenados por magnitud, los promedios se conocen como medidas de tendencia central.

Se definen varios tipos, siendo los más comunes; la media aritmética, la mediana, la moda, la media geométrica y la media armónica. Cada una tiene ventajas y desventajas, según los datos y el objetivo perseguido.

La media aritmética

La media aritmética, o simplemente media, de un conjunto de N números $Y_1, Y_2, ..., Y_N$, se denota por $\bar{Y}$ y se define por:

$$\bar{Y} = \frac{Y_1 + Y_2 + ... + Y_N}{N} = \frac{\sum_{i=1}^{N} Y_i}{N}. \tag{2.1}$$

Si los números $Y_1, Y_2, ..., Y_k$ ocurren, y $f_1, f_2, ..., f_k$ sus frecuencias, respectivas, la media aritmética es

$$\bar{Y} = \frac{f_1Y_1 + f_2Y_2 + ... + f_kY_k}{f_1 + f_2 + ... + f_k} = \frac{\sum_{i=1}^{k} f_iY_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_iY_i}{N}, \tag{2.2}$$

donde $N = \sum_{i=1}^{k} f_i$ es la frecuencia total (es decir, el número total de casos).

La media aritmética ponderada

A veces se asocia a los números $Y_1, Y_2, ..., Y_k$, de ciertos factores de peso (o pesos) $w_1, w_2, ..., w_k$, dependiendo de la influencia asignada a cada número. En tal caso,

$$\bar{Y} = \frac{w_1Y_1 + w_2Y_2 + ... + w_kY_k}{w_1 + w_2 + ... + w_k} = \frac{\sum_{i=1}^{k} w_iY_i}{\sum_{i=1}^{k} w_i}. \tag{2.3}$$

Se le llama media aritmética ponderada con pesos $f_1, f_2, ..., f_k$. Obsérvese la similitud con la ecuación (2.2), que puede considerarse una media aritmética ponderada con pesos $f_1, f_2, ..., f_k$.

Propiedades de la media aritmética

1. La suma algebraica de las desviaciones de un conjunto de números con respecto a su media es cero.

2. La suma de los cuadrados de las desviaciones de un conjunto de números Y_i con respecto de un cierto número a es mínima si y sólo si $a = \bar{Y}$.

3. Si f_1 números tienen media m_1, luego f_2 números tiene media m_2, ..., y hasta f_k números tienen media m_k, entonces la media de todos los números es

$$\bar{Y} = \frac{f_1m_1 + f_2m_2 + ... + f_km_k}{f_1 + f_2 + ... + f_k}, \tag{2.4}$$

es decir, una media aritmética ponderada de todas las medias.

4. Si A es una medida aritmética supuesta o conjeturada y si $d_j = Y_j - A$ son las desviaciones de Y_j respecto de A, las ecuaciones (2.1) y (2.2) se convierten respectivamente en

$$\bar{Y} = A + \frac{\sum_{j=1}^{N} d_j}{N}, \tag{2.5}$$

$$\bar{Y} = A + \frac{\sum_{j=1}^{k} f_j d_j}{\sum_{j=1}^{k} f_j}, \tag{2.6}$$

donde $N = \sum_{j=1}^{k} f_j$. Observe que las fórmulas (2.5) y (2.6) se resumen en la ecuación $\bar{Y} = A + \bar{d}$.

Cálculo de la media aritmética para datos agrupados

Cuando los datos se presentan en una distribución de frecuencias, todos los valores que caen dentro de un intervalo de clase dado, se consideran iguales a la marca de clase, o punto medio del intervalo.

Las fórmulas (2.2) y (2.6) son válidas para datos agrupados y se interpretan como Y_i la marca de clase, f_i como su correspondiente frecuencia de clase, A como cualquier marca de clase conjeturada o supuesta y $d_i = Y_i - A$ como las desviaciones de Y_i respecto de A.

Los cálculos con las fórmulas (2.2) y (2.6) se llaman métodos largos y métodos cortos, respectivamente.

Si todos los intervalos de clase son del mismo tamaño c, las desviaciones $d_i = Y_i - A$ pueden expresarse como cu_i, donde u_i serían números enteros positivos, negativos o cero, es decir, $0, \pm 1, \pm 2, \pm 3, ...$ y la fórmula (1.6) se convierte en

$$\bar{Y} = A + \left(\frac{\sum_{i=1}^{k} f_i u_i}{N} \right) = A + \left(\frac{\sum fu}{N} \right) c, \tag{2.7}$$

que es equivalente a la ecuación $\bar{Y} = A + c\bar{u}$. Esto se conoce como método de codificación para calcular la media.

Es un método corto y debe usarse siempre para datos agrupados con intervalos de clase de tamaños iguales.

Véase que en el método de codificación los valores de la variable Y se transforman en los valores de la variable u de acuerdo con $Y = A + cu$.

La mediana

La mediana de un conjunto de números ordenados en magnitud, es el valor central o la media de los dos valores centrales.

Para datos agrupados, la mediana, obtenida por interpolación, está dada por

$$Mediana = L_1 + \left(\frac{\frac{N}{2} - (\sum f)_1}{f_{mediana}} \right) c, \tag{2.8}$$

donde:

L_1 es la frontera inferior de la clase de la mediana (es decir, la clase que contiene a la mediana).

N es el número total de datos.

$(\sum f)_1$ es la suma de las frecuencias de las clases inferiores a la clase de la mediana.

$f_{mediana}$ es la frecuencia de la clase de la mediana.

c es el tamaño del intervalo de clase de la mediana.

Geométricamente, la mediana es el valor de Y (abscisa), que corresponde a la recta vertical que divide un histograma en dos partes de área igual. Ese valor de Y suele denotarse por $\bar{Y}$.

Moda

La moda de un conjunto de números es el valor que ocurre con mayor frecuencia; es decir, el valor más frecuente. La moda puede no existir e incluso no ser única.

La distribución con una sola moda se llama unimodal.

En el caso de datos agrupados donde se haya construido una curva de frecuencias, para ajustar los datos, la moda será(n) el(los) valor(res) de Y correspondiente(s) al(los) máximo(s) de la curva. Ese valor de Y se denota por $\hat{Y}$.

La moda llega a obtenerse de una distribución de frecuencias o de un histograma a partir de la fórmula:

$$Moda = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c, \tag{2.9}$$

donde:

L_1 es la frontera inferior de la clase modal (clase que contiene a la moda).

Δ_1 es la diferencia de la frecuencia modal con la frecuencia de la clase inferior inmediata.

Δ_2 es la diferencia de la frecuencia modal con la frecuencia de la clase superior inmediata.

c es el tamaño del intervalo de la clase modal.

Relación empírica entre media, mediana y moda

Para curvas de frecuencia unimodales, que sean moderadamente sesgadas o asimétricas, se tiene la siguiente relación empírica:

$$media - moda = 3(media - mediana). \tag{2.10}$$

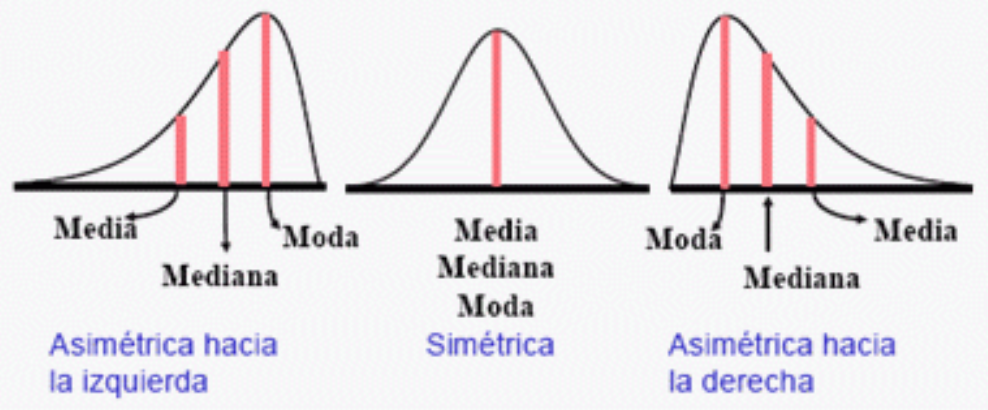

Figura 2.1: Distribución de la media, mediana y moda.

Para curvas simétricas, los valores de la media, la mediana y la moda coinciden.

La media geométrica G

La media geométrica G de un conjunto de N números positivos $Y_1, Y_2, ..., Y_N$ es la raíz N-ésima del producto de esos números:

$$G = \sqrt[N]{Y_1 Y_2 ... Y_N}. \tag{2.11}$$

La media armónica H

La media armónica H de un conjunto de números $Y_1, Y_2, Y_3, ..., Y_N$ es el recíproco de la media aritmética de los recíprocos de los números:

$$H = \frac{1}{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{Y_j}} = \frac{N}{\sum\frac{1}{Y}}. \tag{2.12}$$

En la práctica puede ser más fácil recordar que

$$\frac{1}{H} = \frac{\sum\frac{1}{Y}}{N} = \frac{1}{N}\sum\frac{1}{Y}. \tag{2.13}$$

Relación entre las medias; aritmética, geométrica y armónica

La media geométrica de un conjunto de números positivos $Y_1, Y_2, ..., Y_N$ es menor o igual a su media aritmética, pero es mayor o igual a su media armónica. Es decir,

$$H \leq G \leq \bar{Y}. \tag{2.14}$$

Los signos de igualdad se incluyen sólo si todos los números $Y_1, Y_2, ..., Y_N$ son idénticos.

La media cuadrática (MC)

La media cuadrática (MC) de un conjunto de números $Y_1, Y_2, ..., Y_N$ algunas veces se simboliza por $\sqrt{\bar{Y^2}}$ y se define como

$$MC = \sqrt{\bar{Y^2}} = \sqrt{\frac{\sum_{j=1}^{N} Y_j^2}{N}} = \sqrt{\frac{\sum Y^2}{N}}. \tag{2.15}$$

Este tipo de promedio se utiliza con frecuencia en aplicaciones físicas.

Cuartiles, Deciles y Percentiles

Si un conjunto de datos se ordena de acuerdo con su magnitud, el valor central (o la media aritmética de los dos valores centrales) que divide al conjunto en dos partes iguales es la mediana.

Extendiendo esta idea, es posible considerar los valores que dividen al conjunto en cuatro partes iguales. Estos valores, denotados por Q_1, Q_2 *y* Q_3 se denominan como primero, segundo y tercer cuartíl, respectivamente, donde Q_2 es igual a la mediana.

De forma similar, los valores que dividen en 10 partes iguales son llamados deciles; los cuales se denotan por $D_1, D_2, ..., D_9$, mientras que los valores que dividen a los datos en 100 partes iguales, se conocen como percentiles y se indican con $P_1, P_2, ..., P_{99}$. El quinto decil y el $50o.$ percentil coinciden con la mediana. Los percentiles $25o.$ y $75o.$ corresponden al primero y tercer cuartil, respectivamente.

De manera conjunta, cuartiles, deciles y percentiles, lo mismo que otros valores obtenidos por medio de subdivisiones iguales de los datos, son denominados cuantiles.

2.2. Introducción a la estadística bayesiana

El interés por el teorema de Bayes trasciende a la aplicación clásica, especialmente cuando se amplía a otro contexto en el que la probabilidad no se entiende exclusivamente como la frecuencia relativa de un suceso a largo plazo, sino como el grado de convicción personal acerca de que el suceso ocurra o pueda ocurrir (definición subjetiva de la probabilidad), (Veáse [32]).

Una cuantificación sobre la base subjetiva resulta familiar para el enfoque bayesiano. Al admitir un manejo subjetivo de la probabilidad, el analista bayesiano podrá emitir juicios de probabilidad sobre una hipótesis H y expresar por esa vía su grado de convicción al respecto, tanto antes como después de haber observado los datos.

En su versión más elemental y en este contexto, el teorema de Bayes asume la forma siguiente:

$$Pr(H|datos) = \frac{Pr(datos|H)}{Pr(datos)} Pr(H).$$

La probabilidad a priori de una hipótesis, Pr(H), se ve transformada en una probabilidad a posteriori, Pr(H|datos), una vez incorporada la evidencia que aportan los datos, el uso de probabilidades a priori y a posteriori se veran con más detalles en el siguiente capítulo.

El caso considerado anteriormente se lleva a la situación más simple, aquella en la que Pr(H) representa un número único; sin embargo, si se consiguiera expresar la convicción inicial (y la incertidumbre) mediante una distribución de probabilidades. Entonces una vez observados los datos, el teorema "devuelve" una nueva distribución, que no es otra cosa que la percepción probabilística original actualizada por los datos.

Esta manera de razonar la inferencia bayesiana es radicalmente diferente a la inferencia clásica o frecuentista (que desdeña en lo formal toda información previa de la realidad que examina), es sin embargo, muy cercana al modo de proceder cotidiano, e inductivo. Debe subrayarse que esta metodología, a diferencia del enfoque frecuentista, no tiene como finalidad producir una conclusión dicotómica (significación o no significación, rechazo o no rechazo, etc.) sino que cualquier información empírica, combinada con el conocimiento que ya se tenga del problema que se estudia, "actualiza" dicho conocimiento, y la trascendencia de dicha visión actualizada no depende de una regla mecánica.

Los métodos bayesianos han sido cuestionados argumentando que, al incorporar las creencias o expectativas personales del investigador, pueden ser caldo de cultivo para cualquier manipulación.

Se podría asumir, por una parte que el enfoque frecuentista no está exento de decisiones subjetivas (nivel de significación, usar una o dos colas, importancia que se concede a las diferencias, al establecimiento de pruebas de hipótesis,(veáse [3]); de hecho, la subjetividad (algo muy diferente de la arbitrariedad o el capricho) es un fenómeno inevitable, especialmente en un marco de incertidumbre como en el que operan las ciencias biológicas y sociales. Por otra

parte, las "manipulaciones" son actos de deshonestidad, que pueden producirse en cualquier caso (incluyendo la posibilidad de que se inventen datos) y que no dependen de la metodología empleada sino de la honradez de los investigadores. Aunque las bases de la estadística bayesiana datan de hace más de dos siglos, no es hasta fechas recientes cuando empieza a asistirse en un uso creciente de este enfoque en el ámbito de la investigación. Una de las razones que explican esta realidad y que a la vez anuncian un impetuoso desarrollo futuro es la absoluta necesidad de cálculo computarizado para la resolución de algunos problemas de mediana complejidad.

Hoy ya existen softwares disponible (BUGS, macros para MINITAB, próxima versión de EPIDAT y First Bayes, entre otros) que hace posible operar con estas técnicas y augura el "Inicio de una era Bayesiana".

El proceso intelectual asociado a la inferencia bayesiana es mucho más coherente con el pensamiento usual del científico que el que ofrece el paradigma frecuentista. Los procedimientos bayesianos constituyen una tecnología emergente de procesamiento y análisis de información para la que cabe esperar una presencia cada vez más intensa en el campo de la aplicación de la estadística a la investigación clínica y epidemiológica.

2.2.1. Probabilidad subjetiva

La inferencia bayesiana constituye un enfoque alternativo para el análisis estadístico de datos que contrasta con los métodos convencionales de inferencia, entre otras cosas, por la forma en que asume y maneja la probabilidad, (Veáse [5]). Existen dos definiciones de la noción de probabilidad: objetiva y subjetiva.

La definición frecuentista considera que la probabilidad es el límite de frecuencias relativas (o proporciones) de eventos observables. Pero la probabilidad puede también ser el resultado de una construcción mental del observador, que corresponde al grado de "certeza racional" que tenga acerca de una afirmación, donde el término "probabilidad" significa únicamente que dicha certidumbre está obligada a seguir los axiomas a partir de los que se rige la teoría de probabilidades; como en este marco las probabilidades pueden variar de una persona a otra, se le llaman certidumbres personales, credibilidad, probabilidades personales o grados de creencia.

El teorema de Bayes es el puente para pasar de una probabilidad a priori o inicial, P(H), de una hipótesis H a una probabilidad a posteriori o actualizada, P(H|D), basado en una nueva observación D. Produce una probabilidad conformada a partir de dos componentes: una que con frecuencia se delimita subjetivamente, conocida como "probabilidad a priori", y otra objetiva, la llamada verosimilitud, basada exclusivamente en los datos. A través de la combinación de ambas, el analista conforma entonces un juicio de probabilidad que sintetiza su nuevo grado de convicción al respecto. Esta probabilidad a priori, incorpora la evidencia que aportan los datos, se transforma así en una probabilidad a posteriori.

La regla, axioma o teorema de Bayes, se deriva de manera inmediata a partir de la definición de probabilidad condicional.

Si se tienen dos sucesos A y B (donde A y B son ambos sucesos posibles, es decir, con probabilidad no nula), entonces la probabilidad condicional de A dado B, se define del modo siguiente:

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}, \ \ con\ P(B) \neq 0. \tag{2.16}$$

Análogamente,

$$P(B|A) = \frac{P(B \bigcap A)}{P(A)}, \ \ con\ P(A) \neq 0.$$

De modo que sustituyendo en (2.16) la expresión $P(A \bigcap B) = P(B|A)P(A)$, se llega a la forma más simple de expresar la regla de Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.17}$$

Ahora bien, supóngase que $A_1, A_2, ..., A_k$ son k sucesos mutuamente excluyentes, uno de los cuales ha de ocurrir necesariamente; entonces la conocida (e intuitiva) ley de la probabilidad total establece que:

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i).$$

De modo que, tomando el suceso A_j en lugar de A en la fórmula (2.17) y aplicando al denominador la mencionada ley, se tiene:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}, \ \ j = 1, ..., k, \tag{2.18}$$

que es otra forma en que suele expresarse la regla de Bayes.

La regla de Bayes produce probabilidades inversas, en el sentido de que expresa $P(A|B)$ en términos de $P(B|A)$. La terminología convencional para $P(A|B)$ es la probabilidad a posteriori de A dado B y para $P(A)$ es la probabilidad a priori de A, (debe su nombre a que se aplica antes, sin estar condicionada por la información de que B ocurrió).

2.2.2. ¿Qué es la inferencia bayesiana?

La definición de inferencia estadística en estadística clásica es: la ciencia de extraer conclusiones a partir de una muestra aleatoria para ser aplicadas a cantidades desconocidas de la población de la cuál la muestra fue seleccionada.

Es necesario realizar algunas aclaraciones puntuales respecto de la aproximación clásica con la bayesiana. El punto más importante es que el parámetro, mientras no es conocido, es tratado como una constante en lugar de una variable aleatoria. Esta es la idea fundamental de la teoría clásica pero que conduce a problemas de interpretación.

Por ejemplo, sostener que con 95 % de confianza el intervalo [0.08, 0.12] incluye a la proporción poblacional de los árboles enfermos es incongruente desde que Θ no es aleatorio, Θ está en el intervalo o no lo está. El único elemento aleatorio en este modelo de probabilidad es el dato, por lo tanto, la correcta interpretación del intervalo es que si aplicamos el procedimiento estadístico de construcción de intervalos un gran número de veces, entonces "a la larga" los intervalos construidos incluirán a Θ en el 95 % de dichos intervalos. Todas las inferencias basadas en la teoría clásica son forzadas a tener este tipo de interpretación de frecuencia "a la larga"; a pesar de que en el ejemplo de los árboles, solamente se tiene un intervalo (0.08, 0.12) para realizar el análisis.

El marco teórico en que se aplica la inferencia bayesiana es similar a la clásica: hay un parámetro poblacional respecto al cual se desea realizar inferencias y se tiene un modelo que determina la probabilidad de observar diferentes valores de y, bajo diferentes valores de los parámetros. Sin embargo, la diferencia fundamental es que la inferencia bayesiana considera al parámetro como una variable aleatoria. Esto parecería que no tiene demasiada importancia, pero realmente si lo tiene pues conduce a una aproximación diferente para realizar el modelado del problema y la inferencia propiamente dicha.

En esencia, la inferencia bayesiana esta basada en la distribución de probabilidad del parámetro dado los datos (distribución a posteriori de probabilidad $P(\Theta|y)$, en lugar de la distribución de los datos dado el parámetro. Esta diferencia conduce a inferencias mucho más naturales, lo único que se requiere para el proceso de inferencia bayesiana es la especificación previa de una distribución a priori de probabilidad $P(\Theta)$, la cual representa el conocimiento acerca del parámetro antes de obtener cualquier información respecto a los datos.

La noción de la distribución a priori para el parámetro es el corazón del pensamiento bayesiano. El análisis bayesiano hace uso explícito de las probabilidades para cantidades inciertas (parámetros) en inferencias basadas en análisis estadísticos de datos.

El análisis bayesiano lo podemos dividir en las siguientes etapas:

1. Elección de un modelo de probabilidad completo. Elección de una distribución de probabilidad conjunta para todas las cantidades observables y no observables. El modelo debe ser consistente con el conocimiento acerca del problema fundamental y el proceso de recolección de la información.

2. Condicionamiento de los datos observados. Calcular e interpretar la distribución a posteriori apropiada que se define como la distribución de probabilidad condicional de las cantidades no observadas de interés, dados los datos observados.

3. Evaluación del ajuste del modelo y las implicaciones de la distribución a posteriori resultante. ¿Es el modelo apropiado a los datos?, ¿son las conclusiones razonables?, ¿qué tan sensibles son los resultados a las suposiciones de la modelación de la primera etapa? Si fuese necesario, alterar o ampliar el modelo, y repetir las tres etapas mencionadas.

Existe otro enfoque que permite justificar el método bayesiano, dicho enfoque hace énfasis en las variables observables y en la noción de que un modelo no es más que un artificio probabilístico para hacer predicciones acerca de tales variables, un concepto clave en esta discusión es el de intercambiabilidad, el cual permite, desde una perspectiva subjetivista, justificar el uso (así como clarificar la interpretación) de algunos conceptos estadísticos comunes tales como parámetro,

muestra aleatoria, verosimilitud y distribución inicial.

El supuesto de la "intercambiabilidad", significa que los n valores y_i observados en la muestra pueden ser intercambiados, es decir, que la distribución conjunta $P(y_1, y_2, ..., y_n)$ debe ser invariante a las permutaciones de los índices. Generalmente, los datos de una distribución "intercambiable" es útil modelarlos como independientes e idénticamente distribuidas (iid) dado algún vector de parámetros desconocidos Θ con distribución $p(\Theta)$.

Definición 2.1. *Sea $y_1, y_2, ..., y_n, ...$ una secuencia de cantidades aleatorias. Diremos que son intercambiables bajo una medida de probabilidad P, si para cada subconjunto finito de tal secuencia, su distribución conjunta satisface:*

$$P(y_1, y_2, ..., y_n) = P(y_{\pi(1)}, y_{\pi(2)}, ..., y_{\pi(n)}),$$

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

La definición anterior establece que toda la información relevante, en un conjunto de variables intercambiables, está contenida en los valores de las $y_i's$, de forma que sus índices son "no informativos". Así, el concepto de intercambiabilidad generaliza el de la independencia condicional, así tenemos que, "un conjunto de observaciones independientes e idénticamente distribuidas son siempre un conjunto de observaciones intercambiables".

La inferencia bayesiana se basa en el uso de una distribución de probabilidad para describir todas las cantidades desconocidas relevantes a un problema de estimación, concretamente significa lo siguiente:

Si se dispone de una colección de variables aleatorias intercambiables $\{y_1, y_2, ..., y_n\}$, entonces la distribución de probabilidad

$$f(y_1, y_2, ..., y_n) = \int_\Theta \Pi_{i=1}^n f(y_i|\theta)\pi(\theta)d\theta,$$

donde Θ es la distribución inicial.

$f(y_i|\theta)$ es el modelo de probabilidad.

θ es el límite de alguna función de las observaciones.

$\pi(\theta)$ es una función de probabilidad sobre la distribución inicial Θ.

El concepto de intercambiabilidad es más débil que el de muestra aleatoria simple. Por ejemplo, si las variables intercambiables y_i con $i = 1, 2, ..., n$ toman el valor 0 ó 1, el teorema de representación toma la forma

$$f(y_1, y_2, ..., y_n) = \int_\Theta \Pi_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}\pi(\theta)d\theta,$$

donde: θ=$\lim_{n\to\infty} \frac{\sum_{i=1}^n y_i}{n}$.

Lo cual significa que siempre que se tenga una colección de variables intercambiables, es una muestra aleatoria, si existe una distribución inicial sobre el parámetro θ. El valor del parámetro puede obtenerse como límite de las frecuencias relativas.

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

Teorema 2.1 (Teorema de Representación)**.** *Si* $\{y_1, y_2, ...\}$*, con* $y_i \in \{0, 1\}$*,* $i = 1, ..., n$*, son variables aleatorias intercambiables entonces (ver [4]):*

**Existe* $\theta\epsilon(0, 1)$ *tal que:* $\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n y_i$.

**Existe* $p(\theta)$ *densidad de probabilidad,*

y la densidad conjunta cumple, $P(y_1, ..., y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}p(\theta)d\theta$.

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

Teorema 2.2. *Si* $\{y_i\}$*,* $i\epsilon I$ *es una secuencia infinita de cantidades aleatorias reales intercambiables con medida de probabilidad* P*, entonces, existe una medida de probabilidad* Q *sobre* $\mathfrak{F}$ *tal que la función de distribución conjunta de* $y_1, ..., y_n$ *tiene la forma (ver [5]):*

$$P(y_1, ..., y_n) = \int_{\mathfrak{F}} \prod_{i=1}^n F(y_i)dQ(F),$$

donde,

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

con F_n*-función de distribución empírica definida por* $y_1, ..., y_n$.

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 $p(y)$; en general, para variables aleatorias de cualquier rango y dimensión, la intercambiabilidad identifica las observaciones como una muestra aleatoria de algún modelo probabilístico que garantiza la existencia de una distribución inicial sobre el parámetro que lo describe.

Especificamente, si $y_1, ..., y_n, ...$ es una secuencia intercambiable de cantidades aleatorias reales entonces, existe un modelo paramétrico, $p(y|\theta)$, determinado por algún parámetro $\theta \in \Theta$, el cual es el límite (cuando $n \to \infty$) de alguna función de los $y_i^{'}s$, y existe una distribución de probabilidad para θ, con densidad $p(\theta)$:

$$P(y_1, ..., y_n) = \int_\Theta \prod P(y_i|\theta)P(\theta)d\theta,$$

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 $p(y_i|\theta)$, existiendo una distribución inicial $p(\theta)$, que describe la información inicial sobre el parámetro que determina el modelo.

Por lo tanto, el teorema de la representación, 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.

La aproximación bayesiana implica entonces, que la información muestral y la distribución inicial se actualizan mediante el teorema de Bayes para dar lugar a la distribución final,

$$\pi(\theta|y_1, y_2, ..., y_n) = \frac{\pi(\theta)f(y_1, y_2, ..., y_n|\theta)}{\int_\Theta \pi(\theta)f(y_1, y_2, ..., y_n|\theta)d\theta}.$$

Ahora todas las inferencias, la estimación por punto, la estimación por regiones veraces y los contrastes de hipótesis, se realizan mediante la distribución final.

2.3. El análisis de la robustez y la ayuda a la decisión multicriterio discreta

El término robustez se utiliza con frecuencia en el ámbito de la Ayuda a la Decisión Multicriterio, especialmente en su versión discreta (veáse [8]). De ahí que resulta importante saber qué es la robustez, como se estudia y para que nos sirve.

El objetivo de este trabajo es caracterizar el análisis de la robustez como una forma de respaldar las conclusiones que proporciona el estudio de cualquier problema de decisión multicriterio, y dirimir las dudas en torno a las diversas concepciones del término robustez.

En numerosos estudios e investigaciones en el ámbito de la Ayuda a la Decisión Multicriterio se ha observado la utilización del término robustez. Sin embargo, se suelen atribuir a dicho término

significados muy diferentes e incluso oscuras que pueden conducir a interpretaciones y conclusiones erróneas.

Existen dos grandes líneas de pensamiento en las investigaciones respecto a la Robustez en la Teoría de la Decisión.

Por un lado, la Escuela Franco-Belga con investigadores de la talla de B. Roy, S. Durand, D. Trentesaux, P. Vincke, J.P. Brans, en la cual la robustez aún se encuentra en una fase de análisis de qué es lo que tiene que ser robusto, respecto de qué y por qué debe ser robusto.

Sin embargo, la escuela Anglo-Americana con sus investigadores como J. Rosenhead, J. Mingers, R.L. Ackoff, proponen un análisis de robustez para la presencia de incertidumbre o imprecisión respecto del futuro, tratando de encontrar una medida del grado de robustez a través de un índice numérico.

2.3.1. Tipos de análisis relacionados con el de robustez

Nos parece de gran utilidad comparar y contrastar la noción del análisis de robustez con otros análisis relacionados.

Robustez Estadística

El término "robustez" se utiliza, a menudo en Estadística para hacer referencia a ciertas características deseables de los procesos estadísticos. Se dice que un proceso es robusto respecto de las desviaciones de los supuestos del modelo, cuando el proceso continúa trabajando bien, aún cuando, en mayor o menor extensión, los supuestos no se mantienen. Tales supuestos, a menudo adoptados informáticamente, podrían ser por ejemplo, que una distribución subyacente sea Normal o que las observaciones posean una varianza constante.

En el caso de contrastación de hipótesis estadísticas, un test robusto evita la dificultad de que una decisión (en este caso entre dos hipótesis) se mantenga como muy inestable a la manera de un supuesto particular.

Los bayesianos dan al término un significado más específico. Una aplicación bayesiana es robusta si la distribución posterior de un parámetro desconocido no es significativamente afectada por la elección de la distribución anterior o de la forma del modelo elegido para la generación de los datos. En cualquiera de las dos aproximaciones la incertidumbre, si bien limitada al conocimiento de si los supuestos específicos, en efecto, se mantienen, subyace claramente detrás de la necesidad de disponer y aclarar el concepto en sí mismo. Esto no significa, por supuesto, que debamos estudiar otros tipos de incertidumbre, o la secuencialidad de las decisiones.

Análisis de Sensibilidad

El análisis de sensibilidad es un proceso sistemático utilizado para explorar cómo una solución considerada como "óptima", responde a los cambios introducidos en las condiciones de partida

las cuales típicamente serán valores conocidos que podrán variar en el futuro o parámetros cuyos valores estarán expuestos a cuestionamiento.

De este modo, el análisis se sustenta alrededor del supuesto previo de que la optimización es el escenario principal, con la incertidumbre considerada como un factor potencialmente perjudicial.

El análisis tiene como objetivo estudiar y descubrir el grado de sensibilidad de la solución óptima ante cambios en los factores esenciales. Una solución insensible es considerada como ventajosa y para añadir más confusión lingüística, a veces se le denomina con el término "robusta".

2.3.2. Análisis de robustez desde la perspectiva de la escuela Franco-Belga de ayuda a la decisión multicriterio

De la misma forma que el análisis de sensibilidad, el análisis de la robustez pretende incorporar la experiencia del mundo real respecto de la incertidumbre dentro del entendimiento y la comprensión de los resultados derivados matemáticamente. Podemos señalar dos diferencias importantes con relación al análisis de sensibilidad.

La primera diferencia es que su objetivo consiste, no solamente, en utilizar la optimización sino también un rango más amplio y concreto de resultados informáticos por ejemplo, que una cierta solución sea factible, o que sea casi óptima. La segunda diferencia es que la perspectiva del análisis de robustez se considera virtualmente, como "la imagen reflejada en el espejo" del análisis de sensibilidad. Eso es, identificar cuál es el dominio de puntos en el espacio de soluciones para el cual un resultado particular continúa perpetuándose. La incertidumbre, sin embargo, permanece asociada a los valores de los parámetros, más que al conjunto de incertidumbres intangibles que podrían resistirse a una cuantificación que encierre en sí misma un alto grado de credibilidad.

El propósito de este estudio no consiste en criticar estas formulaciones, sino en distinguir el análisis de robustez de aquellos otros mencionados para clarificar y hacer más reconocibles sus características desde la óptica de la Ayuda a la Decisión Multicriterio. Cada uno de ellos lleva a cabo ciertas funciones que no trata de efectuar el análisis de robustez y viceversa. En primer lugar se debe establecer qué es lo que debe de ser robusto, en segundo lugar respecto de qué, es decir, cómo valorar esa robustez y en tercer lugar la importancia de la robustez en la Teoría de la Decisión Multicriterio. Es lógico pensar que lo que debe ser robusto es el resultado del análisis de decisión efectuado. Entonces para establecer el sujeto de la robustez bastará tener en cuenta el objetivo de la teoría de la decisión multicriterio.

Cualquier método o procedimiento que utilicemos en un planteamiento de Decisión Multicriterio tendrá como objeto calcular una solución al problema propuesto. Parece entonces conveniente que sea la solución robusta y que los analistas que realizan la Ayuda a la Decisión Multicriterio busquen que las soluciones que ellos proponen como resultado de sus procedimientos y algoritmos sean robustas.

En segundo lugar, además de tener definido lo que debe ser robusto, se deben tener claras las razones y factores que pueden dar lugar a la arbitrariedad, la contingencia, la ignorancia, la incertidumbre, respecto de lo que se está analizando, su robustez.

De esta forma, si se pretende dar sentido al término robustez dentro de un contexto decisional específico, resulta importante analizar qué encierra en sí misma cada una de las tres fuentes seña-

ladas precedentemente, como las razones y los factores respecto de los cuales buscamos la robustez.

Este esfuerzo de análisis es necesario no sólo para definir un formalismo adaptado a una modelización apropiada del problema sino también para elegir un método idóneo para su estudio.

En tercer lugar, es necesario preguntarse ¿por qué robustez?.

Así formulada la pregunta parece ser muy amplia, imprecisa y puede dar lugar a respuestas muy variadas. Por ello es conveniente precisarla, limitándola a: por responder a las necesidades, a los tipos de preocupaciones que muestran los demandantes de la ayuda a la decisión. Aquellos decisores que son responsables de adoptar una decisión o bien de influir en un proceso de decisión, no esperan de la ayuda a la decisión más que ella les conduzca, les oriente, o más simplemente, les aporte informaciones útiles para delimitar el campo de reflexión y de acción.

En tales casos, la evaluación del comportamiento puede ser automatizada. De otra forma, se requeriría la discusión entre aquellos que poseen un conocimiento empírico relevante para establecer qué actuaciones o comportamientos son "eficientes" (al menos tan buenos como).

2.3.3. La medida del grado de robustez

Una vez que los procesos de obtención y evaluación del conjunto de alternativas se hayan llevado a cabo, debemos estudiar cuál es el modelo de flexibilidad que ofrecen dichas alternativas, interpretando la flexibilidad como las oportunidades futuras para tomar decisiones orientadas al logro de los objetivos deseados.

Se puede aceptar que la medida de la robustez sea un índice (número real) que tome valores en el intervalo real $[0, 1]$. Una robustez cero indica que existen opciones disponibles aceptables, mientras que una robustez unitaria significa que todas las operaciones disponibles son aceptables.

Así pues, cada conjunto de alternativas o cada alternativa en forma individual tiene un índice de robustez para cada escenario futuro, puesto que una configuración de actuación o comportamiento, variará a lo largo de contextos futuros.

Además, las alternativas pueden ser evaluadas dentro de la extensión de flexibilidad que ofrecen tanto dentro como a través de escenarios futuros. Este proceso, raramente identificará una única alternativa dominante, pero si permitirá eliminar a aquellas no-eficientes y centrar la discusión sobre un número más reducido de alternativas relativamente atractivas y consideradas como buenas soluciones de compromiso.

2.3.4. Otras observaciones del análisis de robustez

Es importante reconocer que el proceso que conduce a la decisión final depende de la identificación de escenarios futuros a los que el sistema en consideración debería afrontar. Esta apreciación es objeto de crítica puesto que el futuro es infinitamente enrevesado o complicado y no podemos conocer, en consecuencia, si alguno de nuestros escenarios futuros identificados podrá

capturar los aspectos claves del escenario actualmente vigente.

Evidentemente, el proceso de obtención y selección preliminar tendería a reducir este riesgo, por ejemplo, seleccionando un amplio rango de posibles estados o entornos futuros contrastables. Sin embargo, esta aproximación no requiere que ese eventual escenario futuro sea identificado actualmente con absoluta certeza.

Así, por ejemplo, si consideramos una alternativa inicial cualquiera, que es el primer paso para llegar a una "solución óptima", en un único escenario futuro previsto, mantendrá una flexibilidad limitada. Por el contrario, una alternativa robusta, mantendrá la flexibilidad a través de un amplio rango de escenarios futuros concebibles.

Lo comentado en líneas anteriores pone en evidencia que la principal ventaja del análisis de robustez subyace más en su proceso que en su producto. No ofrece una simple regla de decisión.

Capítulo 3

Enfoque Bayesiano

En este capitulo definirémos el teorema de Bayes, de una manera más formal, así como el uso de las funciones a priori, los diferentes tipos de estas funciones, las funciones impropias, el teorema central de Límite Bayesiano, aproximación de funciones propias a priori por funciones impropias a priori, para poder culminar con las funciones a posteriori y algunos resultados de funciones a posteriori y Robustes(Ver [18]).

3.1. Conceptos bayesianos básicos

Teorema 3.1. *Sea* $Y = (y_1, y_2, ..., y_n)^{'}$ *un vector de n observaciones cuya distribución de probabilidad* $P(y|\theta)$ *depende de k parámetros involucrados en el vector* $\Theta = (\theta_1, \theta_2, ..., \theta_k)^{'}$. *Supóngase también que* θ *tiene una distribución de probabilidades* $P(\theta)$. *Entonces, la distribución conjunta de* θ *dado Y es:*

$$P(y|\theta) = P(\theta|y)P(\theta),$$

de donde la distribución de probabilidad condicional de θ *dado el vector de observaciones Y resulta:*

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)},$$

con $P(y) \neq 0$.

A esta ecuación se lo conoce como el teorema de Bayes, donde $P(y)$ es la distribución de probabilidad marginal de Y, y puede ser expresada como:

$$P(y) = \begin{cases} \int P(y|\theta)P(\theta)d\theta, & \text{si } \theta \text{ es continuo,} \\ \sum P(y|\theta)P(\theta), & \text{si } \theta \text{ es discreto,} \end{cases}$$

donde la suma o integral es tomada sobre el espacio parámetrico de θ. De este modo, el teorema de Bayes puede ser escrito como:

$$P(\theta|y) = cP(y|\theta)P(\theta) \propto P(y|\theta)P(\theta).$$

Donde:

$P(\theta)$ representa lo que es conocido de θ antes de recolectar los datos y es llamada la distribución a priori de θ;

$P(\theta|y)$ representa lo que se conoce de θ después de recolectar los datos y es llamada la distribución posterior de θ dado Y;

c es una constante normalizadora necesaria para que $P(\theta|y)$ sume o integre uno.

Dado que el vector de datos Y es conocido a través de la muestra, $P(Y|\theta)$ es una función de θ y no de Y. En este caso a $P(Y|\theta)$ se le denomina función de verosimilitud de θ dado Y, y se le denota por $l(\theta|Y)$. Entonces, la fórmula de Bayes puede ser expresada como:

$$P(\theta|y) \propto l(\theta|y)P(\theta).$$

3.1.1. Naturaleza secuencial del teorema de Bayes

Supóngase que se tiene una muestra inicial y_1. Entonces, por la fórmula de Bayes dada anteriormente se tiene:

$$P(\theta|y_1) \propto l(\theta|y_1)P(\theta).$$

Ahora, supóngase que se tiene una segunda muestra y_2 independiente de la primera muestra, entonces:

$$P(\theta|y_1, y_2) \propto l(\theta|y_1, y_2)P(\theta) = l(\theta|y_1)l(\theta|y_2)P(\theta).$$

$$P(\theta|y_1, y_2) \propto l(\theta|y_2)P(\theta|y_1).$$

De esta manera, la distribución a posteriori obtenida con la primera muestra se convierte en la nueva distribución a priori para ser corregida por la segunda muestra. Este proceso puede repetirse indefinidamente. Asi, si se tienen r muestras independientes, la distribución a posteriori

puede ser recalculada secuencialmente para cada muestra de la siguiente manera:

$$P(\theta|y_1, y_2, ..., y_m) \propto l(\theta|y_m)P(\theta|y_1, y_2, ..., y_{m-1}).$$

Nótese que $(\theta|y_1, y_2, ..., y_m)$ podría támbien ser obtenido partiendo de $P(\theta)$ y considerando al total de las r muestras como una sola gran muestra.

La naturaleza secuencial del teorema de Bayes, es tratada por Bernardo (veáse [4]) como un proceso de aprendizaje en términos de probabilidades, el cual permite incorporar al análisis de un problema de decisión, la información proporcionada por los datos experimentales relacionados con los sucesos (parámetros) inciertos relevantes.

3.1.2. Distribución a priori o iniciales

La distribución a priori cumple un papel importante en el análisis Bayesiano ya que mide el grado de conocimiento inicial que se tiene de los parámetros en estudio. Si bien su influencia disminuye a medida que más información muestral es disponible, el uso de una u otra distribución a priori determinará ciertas diferencias en la distribución a posteriori.

Si se tiene un conocimiento previo sobre los parámetros, este se traducirá en una distribución a priori. Así, será posible plantear tantas distribuciones a priori como estados iniciales de conocimiento existan y los diferentes resultados obtenidos en la distribución a posteriori bajo cada uno de los enfoques, adquirirán una importancia en relación con la convicción que tenga el investigador sobre cada estado inicial. Las distribuciones iniciales son parte fundamental de la inferencia Bayesiana. Indudablemente, es el punto más criticado del análisis Bayesiano, la elección de la distribución inicial es 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.

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

$$P(\theta|y) = P(\theta)P(y|\theta).$$

Donde $P(y|\theta)$ el modelo probabilístico, para un conjunto de observaciones dado, $P(\theta)$ es la función inicial. Dadas estas dos funciones se puede calcular la distribución final de θ, es decir, $P(\theta|y)$, la cual es utilizada para hacer inferencias desde el enfoque Bayesiano.

El problema en sí, es el como definir la distribución inicial $P(\theta)$. En esta sección describiremos, la manera en que se puede iniciar.

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 de θ puede ser empleado

para especificar una distribución inicial con una forma funcional particular, entonces una familia paramétrica de funciones puede ser definida (inicial conjugada).

3.1.3. Distribución inicial subjetiva

Consideramos primero la situación en la cual $P(\theta)$ es determinada 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 $P(\theta)$ es considerablemente más difícil.

1. Histograma:

   Cuando Θ es un intervalo sobre la recta real, un enfoque empleado es el 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 $P(\theta)$ suave. 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 trabajar con la densidad obtenida, además, aunque la probabilidad en los extremos de la distribución es frecuentemente pequeña, éstas pueden influenciar las inferencias subsecuentes.

2. Verosimilitud relativa:

   Este enfoque 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:

   $$P(\theta = \theta_0 | \theta = \theta_0 \bigcup \theta = \theta_1) = \frac{P(\theta_0)}{P(\theta_0) + P(\theta_1)},$$

   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 no es acotado, ya que se tiene que determinar la forma de la densidad fuera de una región finita.

3. Adaptación de una forma funcional dada:

   Este enfoque es el más empleado. Consiste en asumir que $P(\theta)$ tiene una forma funcional dada, es decir, $P(\theta) \sim g(\theta|\lambda)$, con $\lambda\ \in\ \Lambda$, entonces es necesario, sólo elejir el hipérparámetro λ.
   Una forma de determinar subjetivamente el parámetro, es calcularlo a partir de los momentos iniciales estimados:

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

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

Desafortunadamente, la estimación de los momentos iniciales, lleva al hecho de que los extremos de la densidad $g(\theta|\lambda)$ pueden llevar a efectos drásticos sobre los momentos, si el parámetro no es 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.

4. Función de distribución acumulativa:

Esta técnica es la menos empleada, aunque razonablemente es más exacta y más fácil de 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 % α percentil (α *cuantil*) de Y si:

$$P(Y \leq z_\alpha) = \alpha, \ con \ \alpha \ \epsilon \ [0, 1].$$

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

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:

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

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 $P(\theta_j|\theta_1, ..., \theta_{j-1})$ para $j = 1, ..., k$, para diferentes valores de $\theta_1, ..., \theta_{j-1}$, así:

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

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).

3.1.4. Distribución a priori no informativa

En ciertos problemas, el conocimiento inicial sobre el verdadero valor del parámetro θ puede ser muy débil, o vago, ésto ha llevado a generar un tipo de distribuciones a priori llamadas distribuciones a priori no informativas, las cuales reflejan un estado de ignorancia inicial.

Una distribución sobre θ se dice que es no informativa si no contiene información sobre θ, es decir, no establece si algunos valores de θ son más favorables que otros.

El postulado de Bayes/Laplace establece que: *cuando nada conocemos acerca de θ en adelante, dada la función a priori $P(\theta)$ es una distribución uniforme, es decir, todos los posibles resultados de θ tienen la misma probabilidad.*

La situación más simple es entonces, cuando se considera 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 $\theta \epsilon \Theta$ la misma densidad, obteniendo así, la inicial no-informativa uniforme $P(\theta) = c$.

El principal problema de emplear la distribución uniforme como inicial no-informativa, es que la distribución uniforme es no invariante bajo reparametrizaciones (veáse [18]), es decir, si no se cuenta con información sobre θ, entonces tampoco sobre $\phi = g(\theta)$ así:

$$P(\theta) = c \Rightarrow \phi = g(\theta) \Rightarrow P(\theta) = \left| \frac{d}{d\theta} g^{-1}(\phi) \right| \cdot P(g^{-1}(\phi)).$$

También, cuando el espacio paramétrico Θ es infinito, la inicial uniforme presenta el problema de ser impropia, aunque no siempre es serio este problema, iniciales impropias, frecuentemente, llevan a posteriores propias (veáse [2, 18]).

Inicial invariante

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 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 (veáse [10]) .

Método de Jeffreys

En situaciones generales, para un parámetro θ el método más usado es el de Jeffreys (1961) que sugiere que, si un investigador es ignorante con respecto a un parámetro θ, entonces su opinión acerca de θ dado las evidencias Y debe ser la misma que el de una parametrización para θ o cualquier transformación uno a uno de θ, $g(\theta)$ una priori invariante sería:

$$P(\theta) \propto \sqrt{I(\theta)},$$

donde $I(\theta)$ es la matriz de información de Fisher:

$$I(\theta) = -E_\theta \left[\frac{\partial^2 \ln f(y|\theta)}{\partial \theta^2} \right].$$

Si θ=$(\theta_1, \theta_2, ..., \theta_n)^{'}$ es un vector, entonces:

$$P(\theta) \propto \sqrt{\det I(\theta)},$$

donde $I(\theta)$ es la matriz de información de Fisher de orden p x p, la posición (ij) de esta matriz es:

$$I_{ij} = -E_0 \left[\frac{\partial^2 \ln f(y|\theta)}{\partial \theta_i \partial \theta_j} \right], \; i, j = 1, ..., n.$$

Por transformación de variables, la densidad a priori $P(\theta)$ es equivalente a la siguiente densidad a priori para ϕ:

$$P(\phi) = P(\theta = h^{-1}(\phi)) \left| \frac{d\theta}{d\phi} \right|.$$

El principio general de Jeffreys consiste en aplicar el método para determinar la densidad a priori $P(\theta)$, debe obtenerse un resultado equivalente en $P(\phi)$ si se aplica la transformación del parámetro para calcular $P(\phi)$ a partir de $P(\theta)$ en la ecuación anterior, o si se obtiene $P(\phi)$ directamente a partir del método inicial. Es decir, debe cumplirse la siguiente igualdad:

$$\sqrt{I(\phi)} = \sqrt{I(\theta)} \left| \frac{d\theta}{d\phi} \right|.$$

Inicial de referencia

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

Definición 3.1. *Se define la información esperada de una observación a partir del modelo* $M = \{p(y|\theta) : y \in Y, \theta \in \Theta\}$*, cuando* $q(\theta)$ *es la distribución inicial para* θ*, como:*

$$I\{q|M\} = \int\int_{Y\times\Theta} p(y|\theta)q(\theta)\log\frac{p(\theta|y)}{q(\theta)}dyd\theta = \int_Y \log\left(\frac{p(\theta|y)}{q(\theta)}\right)p(y)dy,$$

donde $p(\theta|y) = \frac{p(y|\theta)q(\theta)}{p(y)}$*, entonces* $\frac{p(\theta|y)}{q(\theta)} = \frac{p(y|\theta)}{p(y)}$*, esto es por la regla de Bayes,* y *la* $p(y) = \int_\theta p(y|\theta)q(\theta)d\theta$*.*

Considere ahora la información $I\left\{q|M^k\right\}$, la información esperada de k repeticiones independientes de M. Cuando $k \to \infty$, la secuencia de realizaciones $\{y_1, ..., y_k\}$ proporciona toda la información faltante sobre el valor de θ.

Por lo tanto, cuando $k \to \infty$, $I\left\{q|M^k\right\}$ proporciona una medida de la información faltante sobre θ asociada a la inicial $q(\theta)$.

Intuitivamente, una inicial de referencia será una inicial permisible, que maximiza la falta de información sobre θ dentro de la clase $\mathcal{P}$ de iniciales compatibles con cualquier conocimiento asumido sobre el valor θ.

Algunas dificultades que presenta la información esperada es, dentro de los espacios paramétricos continuos, ya que la información faltante $I\left\{q|M^k\right\}$ 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 necesaria la siguiente definición:

Definición 3.2. *Sea* $M = \{p(y|\theta) : y \in Y, \theta \in \Theta \subset \Re\}$*, un modelo paramétrico continuo y sea* $\mathcal{P}$ *una clase de funciones iniciales sobre* θ *para las cuales* $\int_\Theta p(y|\theta)p(\theta)d\theta < \infty$*.*

La función $\pi(\theta)$ se dirá que tiene la propiedad de maximización de la información faltante para el modelo M dado P, si para cualquier conjunto compacto $\Theta_0 \subset \Theta$ y cualquier $p \in \mathcal{P}$:

$$\lim_{k\to\infty}\left\{I\left\{\pi_0|M^k\right\} - I\left\{\mathrm{p}_0|M^k\right\}\right\} \geq 0,$$

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

3.1.5. Distribución a priori conjugada

En este caso, la distribución a priori es determinada completamente por una función de densidad conocida.

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.

Berger (veáse [2]) presenta la siguiente definición para una familia conjugada:

Definición 3.3. *Sea $\mathfrak{F} = \{p(y|\theta) : \theta \in \Theta\}$ una familia de distribuciones muestrales. Una clase $\mathcal{P}$ de distribuciones es llamada una familia conjugada con respecto a $\mathfrak{F}$:*

Si para todo $p(y|\theta) \in \mathfrak{F}$ y $p(\theta) \in \mathcal{P}$ se tiene, $p(\theta|y) \in \mathcal{P}$.

Definición 3.4. *Sea $Y_1, ..., Y_n$ una muestra aleatoria de una población con función de densidad $f(y|\theta)$, una familia D de densidades se dice que es conjugada para la función de densidad $f(y|\theta)$ o que es cerrada bajo muestreo respecto a la función de densidad $f(y|\theta)$, si la función de densidad a priori de θ, $g_\Theta(\theta) \in D$ y si $f_{\theta|Y_1,...,Y_n}(\theta|Y_1, ..., Y_n) \in D$.*

En este caso, la distribución inicial dominará a la función de verosimilitud y $P(\theta|y)$ tendrá la misma forma que $P(\theta)$, con los parámetros corregidos por la información muestral.

Usualmente para una clase de densidades $\mathfrak{F}$, una familia conjugada puede ser determinada examinando la función de verosimilitud $l(\theta|y)$. Entonces la familia conjugada puede ser elegida como la clase de distribuciones con la misma estructura funcional.

A esta clase de familia conjugada se le conoce como distribución a priori conjugada natural.

3.1.6. Funciones impropias

Una propiedad básica de una función de densidad $f(y)$ es que integra (o suma) sobre su rango admisible a 1. Ahora, si $h(y)$ es una función uniforme en la línea real, esto es, $h(y) = c$, si $-\infty < y < \infty$, $c > 0$, entonces es una función impropia, ya que la integral

$$\int_{-\infty}^{\infty} h(y)dy = c\int_{-\infty}^{\infty} dy,$$

no existe, no importa que tan pequeña sea c.

Las funciones de este tipo se emplean frecuentemente para representar el comportamiento local de la distribución a priori en la región donde la verosimilitud es apreciable, pero no sobre su

rango admisible completo. Frecuentemente, una función a priori impropia, se puede combinar con la verosimilitud para dar una densidad posterior que es propia. Esta acción es válida en el sentido de que para un tamaño de muestra grande, la verosimilitud domina la densidad a priori. En este sentido, proporcionan el siguiente teorema Bayesiano.

Teorema 3.2 (Teorema central de Límite Bayesiano [17])**.** *Suponga que $Y_1, Y_2, ..., Y_n$ i.i.d. $f_i(y_i|\theta)$, de tal manera que $f(\underline{y}|\theta) = \prod_{i=1}^{n} f_i(y_i|\theta)$. Suponga además que la densidad a priori $f(\theta)$ y la verosimilitud $f(\underline{y}|\theta)$ son positivas y dos veces diferenciables cerca de $\hat{\theta}$; se asume también que la moda posterior de θ existe. Entonces bajo adecuadas condiciones de regularidad, la distribución posterior $f(\theta|\underline{Y} = \underline{y})$ para n grande, puede aproximarse por una distribución normal con media igual a la moda posterior y matriz de covarianzas igual a menos la inversa de la matriz de segundas derivadas de la log posterior evaluada en la moda.*

Lee (veáse [18]) menciona que es comúnmente razonable analizar datos científicos en la suposición que la verosimilitud domina la densidad a priori por dos razones principales.

Primero, en la ausencia de información a priori, simplemente se debe emplear la función de verosimilitud. Segundo, una investigación científica usualmente no se comienza a menos que la información que proporcione la investigación probablemente sea más precisa que la información disponible; si este es el caso, entonces presumiblemente la verosimilitud dominará la distribución a priori.

3.1.7. Aproximación de función propia a priori por impropias a priori

Teorema 3.3. *Una muestra aleatoria $y = \{y_1, y_2, ..., y_n\}$ de tamaño n es tomada de $N(\theta, \phi)$ donde ϕ es conocida. Supongamos que existen constantes positivas α, ε, M y c dependientes en y (valores muy pequeños de α y ε son de interés), tal que en el intervalo I_α definido por*

$$\bar{y} - \lambda_\alpha \sqrt{(\frac{\phi}{n})} \leq \theta \leq \bar{y} + \lambda_\alpha \sqrt{(\frac{\phi}{n})},$$

donde

$$2\Phi(-\lambda_\alpha) = \alpha,$$

la densidad a priori de θ está entre $c(1-\varepsilon)$ y $c(1+\varepsilon)$, y fuera de I_α es acotado por Mc. Entonces la densidad posterior $p(\theta|y)$ satisface

$$\frac{(1-\varepsilon)}{(1+\varepsilon)(1-\alpha)+M\alpha}(2\pi\frac{\phi}{n})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\bar{y}-\theta)^2/\frac{\phi}{n}\right\} \leq$$

$$p(\theta|y) \leq \frac{(1+\varepsilon)}{(1-\varepsilon)(1-\alpha)}(2\pi\frac{\phi}{n})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\bar{y}-\theta)^2/\frac{\phi}{n}\right\}$$

en el interior de I_α*, y*

$$0 \leq p(\theta|y) \leq \frac{M}{(1-\varepsilon)(1-\alpha)}(2\pi\frac{\phi}{n})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\lambda_\alpha^2\right\}$$

fuera de este intervalo.

3.1.8. Distribución a posteriori

La información a posteriori de θ dado y, con $\theta \in \Theta$, está dada por la expresión $P(y|\theta)$ y expresa lo que es conocido de θ después de observar los datos de y.

Nótese que por la ley multiplicativa de la probabilidad, θ e y tienen la siguiente función de densidad (subjetiva) conjunta,

$$h(y,\theta) = P(\theta)l(\theta|y),$$

en donde:

$P(\theta)$: densidad a priori de θ,

$l(\theta|y)$: función de verosimilitud,

y tiene la función de densidad marginal dada por:

$$m(y) = \int_\Theta l(\theta|y)P(\theta)d\theta.$$

Si la función marginal $m(y) \neq 0$,

$$P(\theta|y) = \frac{h(y,\theta)}{m(y)},$$

sustituyendo

$$P(\theta|y) = \frac{P(\theta)l(\theta|y)}{\int l(\theta|y)P(\theta)d\theta}. \tag{3.1}$$

Que es la misma que la ecuación del teorema de Bayes.

Ahora, dada una muestra de n observaciones independientes, se puede obtener $l(\theta|y)$ y proceder evaluando en la ecuación (3.1). Pero evaluar esta expresión puede ser simplificada, utilizando un estadístico suficiente para θ con función de densidad $g(S(y)|\theta)$, y esto se enuncia en el lema siguiente.

Lema 3.1. *Sea S(y) un estadístico suficiente para θ (es decir, $l(\theta|y) = h(y)g(S(y)|\theta)$), $m(s) \neq 0$ la densidad marginal para $S(y) = s$, entonces se cumple que:*

$$P(\theta|y) = P(\theta|s) = \frac{g(s|\theta)P(\theta)}{m(s)}.$$

Demostración:

$$P(\theta|y) = \frac{l(\theta|y)P(\theta)}{\int l(y|\theta)P(\theta)d\theta}$$

$$= \frac{h(y)g(S(y)|\theta)P(\theta)}{\int h(y)g(S(y)|\theta)P(\theta)d\theta}$$

$$= \frac{g(s|\theta)P(\theta)}{m(s)} = P(\theta|s). \ \blacksquare$$

Se observa que la ecuación (1.4) se puede expresar de la manera más corta conveniente,

$$P(\theta|y) \propto l(y|\theta)P(\theta). \tag{3.2}$$

En otras palabras, la distribución a posteriori es proporcional a la verosimilitud por la a priori, es decir, que la probabilidad es multiplicada por una constante (o una función de y), sin alterar a la a posteriori, pués en la ecuación (1.4) el denominador no depende de θ. En la práctica, el cálculo de la distribución a posteriori puede ser un asunto complicado, especialmente si la dimensión del parámetro no es pequeña. Sin embargo, para ciertas combinaciones de distribuciones iniciales y verosimilitudes es posible simplificar el análisis. En otros casos se requieren aproximaciones analíticas y/o técnicas computacionales relativamente sofisticadas.

Función posterior de una distribución normal a priori

Decimos que y tiene una distribución normal de media θ y varianza ϕ, la escribimos

$$y \sim N(\theta, \phi),$$

donde,

$$p(y) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\theta)^2/\phi\right\}.$$

Supongamos que ϕ conocido y θ desconocido, puede ser expresado en términos de una normal,

$$\theta \sim N(\theta_0, \phi_0).$$

Consideremos una observación $y \sim N(\theta, \phi)$ con media igual al parámetro de interés, donde θ_0, ϕ_0 son conocidos.

$$p(\theta) = (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/\phi_0\right\},$$

$$p(y|\theta) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \theta)^2/\phi\right\}.$$

Por lo tanto,

$$p(\theta|y) = p(\theta)p(y|\theta),$$

$$p(\theta|y) = (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/\phi_0\right\} \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \theta)^2/\phi\right\}.$$

Aplicando la regla de Bayes nos da:

$$p(\theta|y) \propto \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/\phi_0\right\} \times \exp\left\{-\frac{1}{2}(y - \theta)^2/\phi\right\}.$$

Desarrollando los cuadrados y haciendo los cálculos correspondientes requeridos y aplicando nuevamente la regla de Bayes tenemos lo siguiente:

$$p(\theta|y) \propto \exp\left\{-\frac{1}{2}\theta^2(\phi_0^{-1} + \phi^{-1}) + \theta(\theta_0/\phi_0 + y/\phi)\right\},$$

recordando que $p(\theta|y)$ como una función de θ.

Ahora, esto nos conviene escribir de esta manera:

$$\phi_1 = \frac{1}{\phi_0^{-1} + \phi^{-1}},$$

$$\theta = \phi_1(\theta_0/\phi_0 + y/\phi),$$

tal que

$$\phi_0^{-1} + \phi^{-1} = \phi_1^{-1}.$$

A este término se le conoce como la precisión, que es definida como el recíproco de la varianza. Puede ser recordado como:

$$\text{precisión posterior} = \text{precisión a priori} + \text{precisión de los datos},$$

y por lo tanto,

$$p(\theta|y) \propto exp\left\{-\frac{1}{2}\theta^2/\phi + \theta\theta_1/\phi\right\}.$$

Agregando dentro del exponencial el término

$$-\frac{1}{2}\theta_1^2/\phi_1.$$

Vemos que

$$p(\theta|y) \propto \exp\left\{-\frac{1}{2}(\theta-\theta_1)^2/\phi\right\}.$$

De esto tenenos que

$$p(\theta|y) \propto (2\pi\phi_1)^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\theta-\theta_1)^2/\phi\right\}.$$

Esto nos da que la densidad posterior es

$$\theta|y \sim N(\theta_1, \phi_1).$$

La relación para la media posterior, θ_1 es sólo un poco más complicado.

Tenemos

$$\theta_1 = \theta_0 \frac{\phi_0^{-1}}{\phi_0^{-1}+\phi^{-1}} + \frac{\phi^{-1}}{\phi_0^{-1}+\phi^{-1}}.$$

Distribución predictiva

El caso discutido en esta sección, es fácil de encontrar la distribución predictiva, ya que:

$$\bar{y} = (\bar{y}-\theta) + \theta,$$

y se tiene,

$$\bar{y} - \theta \sim N(0, \phi),$$
$$\theta \sim N(\theta_0, \phi_0),$$

de esto tenemos que,

$$\bar{y} \sim N(\theta_0, \phi + \phi_0).$$

Usando que la suma de varias normales independientes tiene una distribución normal. Con media igual a la suma de las medias y varianza igual a la suma de las varianzas.

Distribución posterior de una normal multivariada

Podemos generalizar la situación previa, supongamos una función a priori

$$\theta \sim N(\theta_0, \phi_0).$$

Supongamos ahora que tenemos n observaciones independientes $y = (y_1, y_2, ..., y_n)$ tal que

$$y_i \sim N(\theta, \phi),$$

$$p(\theta|y) \propto p(\theta)p(y|\theta) = p(\theta)p(y_1|\theta)p(y_2|\theta) \ ... \ p(y_n|\theta)$$

$$= (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/\phi_0\right\}$$
$$\times(2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_1 - \theta)^2/\phi\right\}$$
$$\times(2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_2 - \theta)^2/\phi\right\}$$
$$\times ... \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_n - \theta)^2/\phi\right\}$$
$$\propto \exp\left\{-\frac{1}{2}\theta^2(\frac{1}{\phi_0} + \frac{n}{\phi}) + \theta(\frac{\theta_0}{\phi_0} + \frac{\sum_{i=1}^n y_i}{\phi})\right\}.$$

Procediendo como cuando tenemos una observación, se da lo siguiente:

$$\theta \sim N(\theta_1, \phi_1).$$

Donde

$$\phi_1 = (\frac{1}{\phi_0} + \frac{n}{\phi})^{-1}$$
$$\theta_1 = \phi_1 \left\{ \frac{\theta_0}{\phi_0} + \frac{\bar{y}}{\frac{\phi}{n}} \right\}.$$

Asumiendo una normal a priori y máxima verosimilitud, el resultado es justo el mismo como la distribución posterior obtenida de una observación singular de la media $\bar{y}$, por lo tanto, sabemos que

$$\bar{y} \sim N(\theta, \frac{\phi}{n}).$$

Distribución predictiva

Si consideramos tomar una simple observación y_{n+1}, entonces la distribución predictiva, puede ser encontrada justo como en el caso anterior [18]:

$$y_{n+1} = (y_{n+1} - \theta) + \theta$$

y más que nada, independientemente de una u otra,

$$(y_{n+1} - \theta) \sim N(0, \phi),$$
$$\theta \sim N(\theta_1, \phi_1),$$

tal que

$$y_{n+1} \sim N(\theta_1, \phi + \phi_1).$$

Es simple adaptar este argumento para encontrar la distibución predictiva de un vector de m valores $y = (y_1, y_2, ..., y_m)$ donde

$$y_i \sim N(\theta, \psi),$$

por escribir

$$y = (y - \theta * I) + \theta * I.$$

Donde I es el vector constante

$$I = (1, 1, ..., 1).$$

Entonces θ tiene su distribución posterior $\theta|x$ y los componentes del vector $y - \theta * I$ son $N(0, \psi)$ variables independientes de θ y de las variables aleatorias y y x, tal que y tiene una distribución normal multivariada, además sus componentes no son independientes unos de otros (veáse [18]).

Una conveniente a priori para la varianza de una normal

Supongamos que tenemos una muestra simple de n observaciones $y = (y_1, y_2, ..., y_n)$ de $N(\mu, \phi)$ donde la varianza ϕ es desconocida, pero la media μ es conocida. Entonces

$$\begin{aligned} p(y|\phi) &= (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_1 - \mu)^2/\phi\right\} \\ &\times ... \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_n - \mu)^2/\phi\right\} \\ &\propto \phi^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2/\phi\right\}. \end{aligned}$$

Al tomar

$$S = \sum_{i=1}^{n}(y_i - \mu)^2.$$

Recordamos que μ es conocida; sustituimos lo anterior en la parte de arriba por conveniencia.

$$p(y|\phi) \propto \phi^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}S/\phi\right\}.$$

En principio, simplemente tenemos una forma de distribución a priori para la varianza ϕ. De cualquier modo, si somos capaces de encontrar la distribución posterior, nos puede servir mas si la distribución posterior es adecuada. Esto pasa si la a priori se asemeja a la verosimilitud,

$$p(\phi) \propto \phi^{-\frac{k}{2}} \exp\left\{-\frac{1}{2}S_0/\phi\right\},$$

donde k y S_0 son constantes convenientes. Tratando de encontrar una familia de distribuciones adecuada se escribe $k = v + 2$ tal que

$$p(\phi) \propto \phi^{-\frac{v+2}{2}} \exp\left\{-\frac{1}{2}S_0/\phi\right\}.$$

Se tiene primero para la distribución posterior

$$\begin{aligned} p(\phi|y) &\propto p(\phi)p(y|\phi) \\ &\propto \phi^{-\frac{v+n}{2}-1} \exp\left\{-\frac{1}{2}(S_0+S)/\phi\right\}. \end{aligned}$$

Esta distribución pertenece a la distribución chi-cuadrada o χ^2. Esto lo vemos más claro si trabajamos en términos de la precisión $\lambda = \frac{1}{\phi}$ en lugar del término de la varianza ϕ, por lo tanto

$$\begin{aligned} p(\lambda|y) &\propto p(\phi|y)|d\phi| \\ &\propto \phi^{-\frac{v+n}{2}+1} \exp\left\{-\frac{1}{2}(S_0+S)\lambda\right\} \times \lambda^{-2} \\ &\propto \phi^{-\frac{v+n}{2}-1} \exp\left\{-\frac{1}{2}(S_0+S)\lambda\right\}. \end{aligned}$$

La expresión se parece a una distribución chi-cuadrada.

De las propiedades de esta distribución se tiene:

$$(S_0 + S)\lambda \sim \chi^2_{v+n}.$$

De donde $(S_0 + S)\lambda$ tiene una distribución χ^2 con $v + n$ grados de libertad.

$$\phi \sim (S_0 + S)\chi^{-2}_{v+n}.$$

Y decimos que ϕ tiene una distribución Chi-cuadrada inversa. Si

$$\phi \sim S_0 \chi_v^{-2}.$$

De esto se tiene:

$$E(\phi) = S_0/(v-2), \;\; V(\phi) = \frac{2S_0^2}{(v-2)^2(v-4)}.$$

Normal con media y varianza desconocida

Supongamos que ambos parámetros de la distribución normal son desconocidos. Consideremos el caso donde tenemos un conjunto de observaciones $y = (y_1, y_2, ..., y_n)$ donde $N(\theta, \phi)$ con θ y ϕ ambos desconocidos.

$$p(y|\theta,\phi) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\theta)^2/\phi\right\}$$
$$= \{2\pi\}^{-\frac{1}{2}}\{\phi\}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\theta^2/\phi\right\} \exp\left\{-\frac{1}{2}y\theta/\phi - \frac{1}{2}y^2\right\}.$$

Utilizando el hecho de la definición de la familia exponencial, en el siguiente capítulo se definirá con más detalle dicho concepto,

$$l(\theta,\phi|y) \propto p(y|\theta,\phi)$$
$$= (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_1-\theta)^2/\theta\right\}$$
$$\times ... \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_n-\theta)^2/\theta\right\}$$
$$\propto \phi^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2/\phi\right]$$
$$= \phi^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\theta)^2\right\}/\phi\right]$$
$$= \phi^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\left\{S + n(\bar{y}-\theta)^2\right\}/\phi\right].$$

Donde definimos

$$S = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Y definiendo a

$$s^2 = S/(n-1).$$

Sustituyendo a $s^2 = S/n$ donde la media es conocida, para facilitar los cálculos, sí poder determinar a que tipo de distribución tiende.

Ahora pasamos a trabajar el espacio vectorial (ver [18]) de dos dimensiones $(\bar{y}, S)$ o equivalentemente $(\bar{y}, s^2)$ es claramente suficiente para (θ, ϕ) dado y.

Porque este caso puede ser más complicado, primero consideremos el caso de la a priori. Es usual tomar

$$p(\theta, \phi) \propto 1/\phi.$$

Con esto, el producto para la referencia a priori $p(\theta) \propto 1$ para θ y la referencia a priori $p(\phi) \propto 1/\phi$ para ϕ. La justificación para esto es que parece improbable que si sabemos acerca de la media o la varianza, entonces nos da información acerca de lo que podría afectar una de la otra,

$$p(\theta, \phi|y) \propto \phi^{-\frac{n}{2}-1} \exp\left[-\frac{1}{2}\left\{S + n(\bar{y}-\theta)^2\right\}/\phi\right].$$

Por esta razón nos conviene escribir [18]

$$v = n - 1,$$

en la función de θ, no en la exponencial.

$$p(\theta, \phi|y) \propto \phi^{-\frac{v+1}{2}-1} \exp\left[-\frac{1}{2}\left\{S + n(\bar{y}-\theta)^2\right\}/\phi\right].$$

Distribución posterior de una distribución binomial

Sea el parámetro θ que tiene una distribución a priori uniforme en el intervalo [0,1] y la variable aleatoria Y que tiene una distribución de probabilidades Binomial con parámetros m y θ, m conocido por conveniencia. Entonces se tienen las siguientes funciones de distribución

$$p(\theta) = 1, \;\; 0 \leq \theta \leq 1.$$

$$p(y|\theta) = \begin{cases} \binom{m}{y}\theta^y(1-\theta)^{m-y}, & y = 0, 1, ...m. \\ 0, & y \neq 0, 1, ..., m. \end{cases}$$

Ahora, para una muestra aleatoria de tamaño n la función de verosimilitud estará dada por

$$l(\theta|y) = [\prod_{i=1}^{n} \binom{m}{y}]\theta^{\sum_{i=1}^n y_i}(1-\theta)^{nm-\sum_{i=1}^n y_i}, \;\; y_i = 0, 1, ...m.$$

y aplicando el Teorema de Bayes, la distribución a posteriori de θ dada la muestra, queda expresada como:

$$p(\theta|y) = c\frac{n(m!)}{\prod_{i=1}^n (y_i)! \prod_{i=1}^n (m-y_i)!}\theta^{\sum_{i=1}^n y_i}(1-\theta)^{mn-\sum_{i=1}^n y_i}.$$

Esta expresión puede escribirse de la siguiente manera,

$$p(\theta|y) = c\frac{n(m!)}{\prod_{i=1}^n (y_i)! \prod_{i=1}^n (m-y_i)!}\theta^{\sum_{i=1}^n (y_i+1)-1}(1-\theta)^{(mn-\sum_{i=1}^n (y_i+1))-1}$$

que tiene la forma de una distribución Beta con parámetros $(\sum_{i=1}^n (y_i+1))$ y $(nm-\sum_{i=1}^n (y_i+1))$,

$$\theta \sim Be\left(\sum_{i=1}^{n}(y_i+1), nm - \sum_{i=1}^{n}(y_i+1)\right).$$

Luego el valor adecuado de la constante normalizadora c será:

$$c = \frac{\Gamma(nm+2)}{\Gamma(\sum_{i=1}^n (y_i+1))\Gamma(nm-\sum_{i=1}^n (y_i+1))}\frac{\prod_{i=1}^n y_i! \prod_{i=1}^n (m-y_i)!}{n(m!)}$$

Nótese que es a través de $l(\theta|y)$ que los datos (información muestral) modifican el conocimiento previo de θ dado por $p(\theta)$.

3.2. Robustez

Note que cualquier declaración de una distribución a posteriori y cualquier inferencia es condicional, no simplemente en los datos, y también en la suposición hecha acerca de la máxima verosimilitud. Decimos que una inferencia es robusta si esta no es seriamente afectada por cambios en la suposición en que esta se basa (veáse [14]).

Capítulo 4

Inferencia Bayesiana

Los problemas concernientes a la inferencia de θ pueden ser resueltos fácilmente utilizando análisis bayesiano. Dado que la distribución a posteriori contiene toda la información disponible acerca del parámetro, muchas inferencias concernientes a θ pueden consistir únicamente de las características de esta distribución. Es por eso, que en éste capítulo, se expondrá los diferentes tipos de estimación bayesiana, necesarios para el análisis bayesiano, intervalos de credibilidad y pruebas de hipótesis (veáse [18, 32]).

4.1. Estimación puntual por máxima verosimilitud

Para estimar θ, se pueden aplicar numerosas técnicas de la estadística clásica a la distribución a posteriori. La técnica más conocida es la estimación por máxima verosimilitud, en la cual se elige como estimador de θ, al valor $\hat{\theta}$ que maximiza a la función de verosimilitud $l(\theta|y)$ (ver [11]).

Análogamente, la estimación bayesiana por máxima verosimilitud se define de la manera siguiente:

Definición 4.1. *El estimador generalizado de máxima verosimilitud de θ es definido como la moda más grande, $\hat{\theta}$ de $\pi(\theta|y)$. En otras palabras el valor de $\hat{\theta}$ que maximiza a $\pi(\theta|y)$, es considerado como función de θ.*

4.1.1. Error de estimación

Cuando se hace una estimación, es usualmente necesario indicar la precisión de la estimación. La medida bayesiana que se utiliza para medir la precisión de una estimación (en una dimensión) es la varianza a posteriori de la estimación (ver [11]).

Definición 4.2. *Si θ es un parámetro de valor real con distribución a posteriori $\pi(\theta|y)$, y δ es el estimador de θ, entonces la varianza a posteriori de δ es*

$$V_\delta^\pi(y) = E^{\pi(\theta|y)}[(\theta - \delta)^2].$$

Nótese en la definición anterior, al tener el estimador $\hat{\theta}$ de θ, en este caso δ, solo se necesita sustituirla en la definición de la varianza para obtener la varianza a posteriori.

Cuando δ es la media a posteriori,

$$\mu^\pi(y) = E^{\pi(\theta|y)}[\theta],$$

entonces $V^\pi(y) = V_\theta^\pi \pi(y)$ será llamada varianza a posteriori (la varianza de θ para la distribución $\pi(\theta|y)$).

La desviación estándar a posteriori es $\sqrt{V^\pi(y)}$. Generalmente se utiliza la desviación estándar a posteriori $\sqrt{V_\delta^\pi(y)}$ del estimador δ, como el error estándar de la estimación δ.

Para simplificar cálculos que serán utilizados más adelante, se puede representar a la varianza a posteriori de la siguiente forma:

$$\begin{aligned} V_\delta^\pi(y) &= E^{\pi(\theta|y)}[(\theta - \delta)^2] \\ &= E[(\theta - \mu^\pi(y) + \mu^\pi(y) - \delta)^2] \\ &= E[(\theta - \mu^\pi(y))^2] + E[2(\theta - \mu^\pi(y))(\mu^\pi(y) - \delta)] + E[(\mu^\pi(y) - \delta)^2] \\ &= V^\pi(y) + 2(\mu^\pi(y) - \delta)(E[\theta] - \mu^\pi(y)) + (\mu^\pi - \delta)^2 \\ &= V^\pi(y) + (\mu^\pi - \delta)^2. \end{aligned}$$

Así, entonces

$$V_\delta^\pi(y) = V^\pi(y) + (\mu^\pi - \delta)^2. \tag{4.1}$$

4.1.2. Estimación multivariada

La estimación Bayesiana de un vector $\theta = (\theta_1, ..., \theta_n)$ es la estimación de máxima verosimilitud generalizada (moda a posteriori) $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_n)$. sin embargo, por cuestiones de cálculo y porque la unicidad y existencia no se garantizan en el caso multivariado, es más conveniente utilizar la medida a posteriori (ver [11])

$$\mu^\pi(y) = (\mu_1^\pi(y), \mu_2^\pi(y), ..., \mu_n^\pi(y))^t = E^{\pi(\theta|y)}[\theta],$$

como el estimador bayesiano, y la precisión puede ser descrita por la matriz de covarianza a posteriori.

$$V^{\pi}(y) = E^{\pi(\theta|y)}[\theta - \mu^{\pi}(y)]^t.$$

El error estándar de la estimación $\mu_i^{\pi}(y)$ de $\hat{\theta}_i$ sería $\sqrt{V_{ii}^{\pi}(y)}$, en donde $V_{ii}^{\pi}(y)$ es el elemento (i,i) de $V^{\pi}(y)$. Análogamente a la ecuación (3.1) se puede escribir:

$$\begin{aligned} V_{\delta}^{\pi}(y) &= E^{\pi(\theta|y)}[(\theta - \delta)(\theta - \delta)^t] \\ &= V^{\pi}(y) + (\mu^{\pi}(y) - \delta)(\mu^{\pi}(y) - \delta)^t. \end{aligned}$$

4.2. 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 θ, que rige el modelo, estará contenida en la distribución posterior $p(\theta|D)$.

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 (ver [2]).

Es conveniente resumir la información contenida en la distribución posterior por:

i) Proponer valores para la cantidad de interés, que conforme a los resultados de los datos, que son verosímiles a estar más cerca al verdadero valor.

ii) Medir la compatibilidad de los resultados con los valores hipotéticos de la cantidad de interés, la cual puede ser sugerida por el contexto de la investigación.

La distribución posterior reemplaza la función de verosimilitud como una expresión que incorpora toda la información. $\prod(\theta|y)$ es un resumen completo de la información acerca del parámetro θ.

Sin embargo, para algunas aplicaciones es deseable (o necesario) resumir esta información en alguna forma.

Especialmente, si se desea proporcionar un simple "mejor" estimado del parámetro desconocido. (Nótese la distinción con la estadística clásica en que los estimadores puntuales de los parámetros son la consecuencia natural de una inferencia).

Por lo tanto, en el contexto bayesiano, ¿cómo se puede reducir la información en una $P(\theta|y)$ a un "mejor" estimado?, ¿qué se debe entender por "mejor"? (ver [32]).

Existen dos formas de enfrentar el problema:

a) Realizar estimación de Bayes posterior.

b) Aproximación de teoría de decisión.

4.2.1. Estimador de Bayes posterior

El estimador de Bayes posterior (ver [4]) se define de la siguiente manera:

Definición 4.3. *Sean $y_1, y_2, ..., y_n$ una muestra aleatoria de una población con función de densidad $f(y|\theta)$, $g_\Theta(\theta)$ la función de densidad a priori de Θ y $\pi(\theta)$ una función del parámetro θ. El estimador bayesiano para la imagen de θ bajo la función $\pi(\theta)$, respecto a la función de densidad a priori $g_\Theta(\theta)$ es aquel cuya estimación corresponde a*

$$E(\pi(\theta)|(y_1, y_2, ..., y_n)) = \frac{\int_{-\infty}^{\infty} \pi(\theta)[\prod_{i=1}^n f(y_i|\theta)]g_\Theta(\theta)d\theta}{\int_{-\infty}^{\infty}[\prod_{i=1}^n f(y_i|\theta)]g_\Theta(\theta)d\theta}.$$

Una vez que ya hemos calculado $E(\pi(\theta)|(y_1, ..., y_n))$, encontramos el estimador bayesiano para θ, minimizando alguna función de pérdida. Es decir tenemos que:

$$\hat{\theta}_{Bayes} = argmin\ E(L(\theta, \delta(y_1, ..., y_n))).$$

En el siguiente capítulo, se describe de una manera más formal lo que es un función de pérdida [9], pero antes, una aplicación del teorema anterior.

Aplicación del estimador de Bayes:

Sean $\{y_1, y_2, ..., y_n\}$ una muestra aleatoria de $f(y|\theta) = \theta^y(1-\theta)^{1-y}$ para y=1,0 y $g_\theta(\theta) = I_{(0,1)}(\theta)$. ¿Cuáles son los estimadores de θ y $\theta(1-\theta)$?.

$$f(\theta|y_1, y_2, ..., y_n) = \frac{g_\theta(\theta)\prod_{i=1}^n f(y_i|\theta)}{\int_0^1 g_\theta(\theta)\prod_{i=1}^n f(y_i|\theta)d\theta}$$

$$f(\theta|y_1, y_2, ..., y_n) = \frac{\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} I_{(0,1)}(\theta)}{\int_0^1 \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} d\theta}$$

$$E(\theta|y_1, y_2, ..., y_n) = \frac{\int_0^1 \theta \left(\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}\right) d\theta}{\int_0^1 \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} d\theta}$$

$$E(\theta|y_1, y_2, ..., y_n) = \frac{B\left(\sum_{i=1}^n y_i + 2,\ n - \sum_{i=1}^n y_i + 1\right)}{B\left(\sum_{i=1}^n y_i + 1,\ n - \sum_{i=1}^n y_i + 1\right)}$$

$$E(\theta|y_1, y_2, ..., y_n) = \frac{\sum_{i=1}^n y_i + 1}{n+2}.$$

Luego el estimador a posteriori de Bayes de θ es $\frac{\sum_{i=1}^n y_i + 1}{n+2}$, el cual es un estimador sesgado. El estimador máximo verosímil de θ es $\frac{\sum_{i=1}^n y_i}{n}$, que es un estimador insesgado. Además,

$$E(\theta(1-\theta)|y_1, y_2, ..., y_n) = \frac{\int_0^1 \theta(1-\theta)\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} d\theta}{\int_0^1 \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} d\theta}$$

$$E(\theta(1-\theta)|y_1, y_2, ..., y_n) = \frac{\Gamma(\sum_{i=1}^n y_i + 2)\Gamma(n - \sum_{i=1}^n y_i + 2)}{\Gamma(n+4)}$$

$$E(\theta(1-\theta)|y_1, y_2, ..., y_n) = \frac{\Gamma(n+2)}{\Gamma(\sum_{i=1}^n y_i + 1)\Gamma(n - \sum_{i=1}^n y_i + 1)}$$

$$E(\theta(1-\theta)|y_1, y_2, ..., y_n) = \frac{(\sum_{i=1}^n y_i + 1)(n - \sum_{i=1}^n y_i + 1)}{(n+3)(n+2)},$$

el cual es un estimador de $\theta(1-\theta)$ con respecto a la apriori uniforme.

4.2.2. Aproximación por teoría de decisión

Teoría de decisión bayesiana: consiste en el análisis de las principales alternativas posibles. Diferenciar aquellas circunstancias que pueden producir un resultado distinto. Evaluar y analizar las posibilidades de que se den distintas circunstancias. Calcular el valor esperado para cada decisión y elegir la decisión con el valor mas elevado (ver [2]).

En la teoría de decisión la idea es combinar la información muestral con información no muestral con el objeto de tomar una decisión óptima. El análisis bayesiano comparte con la teoría de la decisión el uso de información no muestral (ver [2]).

Recordemos que Θ es el espacio de parámetros o estados de la naturaleza, $\theta \in \Theta$ es un estado de la naturaleza.

Sea A el espacio de acciones del tomador de decisiones, $a \in A$ es una acción.

Un problema de decisión es una función $D : A \times \Theta \rightarrow C$, donde C es un espacio de consecuencias. Supongamos que el agente tiene preferencias sobre el conjunto de consecuencias que las representamos mediante una función de utilidad o de pérdida.

A continuación definimos la función de pérdida como la composición de la función D y la función de utilidad o de pérdida.

Un problema de desición está bien puesto cuando el conjunto de acciones, estados de la naturaleza y consecuencias son tales preferencias del tomador de decisiones, sobre las consecuencias son totalmente independientes de las acciones o estados de la naturaleza.

Sea $L(\theta, a)$ una función de pérdida.

Definimos la pérdida esperada Bayesiana cuando se toma una decisión $a \in A$ como:

$$p(a|y) = \int_{\Theta} L(\theta, a) f(\theta|y) d\theta.$$

Dada una función de pérdida y una distribución posterior definimos al estimador bayesiano de θ como [9]:

$$\hat{\theta}_{Bayes}(y) = argmin\ p(a|y).$$

Funciones de pérdida

Se especifica una función de pérdida $L(\theta, a)$ que cuantifíca las posibles penalidades en estimar θ por a.

Hay muchas funciones de pérdida que se puede usar, la elección en particular de una de ellas dependerá del contexto del problema.

Las más usadas son:

1. Pérdida cuadrática: $L(\theta, a) = (\theta - a)^2$.

2. Pérdida del error absóluto: $L(\theta, a) = |\theta - a|$.

3. Pérdida 0, 1: $L(\theta, a) = \begin{cases} 1, & si\ |a - \theta| \leq \epsilon, \\ 0, & si\ |a - \theta| > \epsilon. \end{cases}$

4. Pérdida lineal: para $g, h > 0$

$$L(\theta, a) = \begin{cases} g(a-\theta), & si\ a > \theta, \\ h(\theta - a), & si\ a < \theta. \end{cases}$$

En cada uno de los casos anteriores, por la minimización de la pérdida esperada posterior, se obtienen formas simples para la regla de decisión de Bayes, que es considerado como el estimador puntual de θ para la elección en particular de la función de pérdida (ver [2]).

Notas:

- $L(\theta, a)$ es la pérdida incurrida en adoptar la acción a cuando el verdadero estado de la naturaleza es θ.

- $p(a, y)$ es la pérdida posterior.

Luego $R_a = E(L(\theta, a)) = p(a, y) = \int L(\theta, a)p(\theta/y)d\theta$.

- Regla de decisión de Bayes (estimador de Bayes): $d(y)$ es la acción que minimiza a $p(a, y)$.

- Riesgo de Bayes es $BR(d) = \int p(d(y), y)p(y)dy$.

En el contexto bayesiano, un estimador puntual de un parámetro es una simple estadística descriptiva de la distribución posterior $\prod(\theta|y)$ (ver [2]).

Utilizando la calidad de un estimador a través de la función de pérdida, la metodología de la teoría de decisión conduce a elecciones óptimas de estimados puntuales.

En particular, las elecciones más naturales de función de pérdida conducen respectivamente a la media posterior, mediana y moda como estimadores puntuales óptimos.

Una aplicación

Sean $\{y_1, y_2, ..., y_n\}$ una muestra aleatoria de una distribución normal, $N(\theta, 1)$, $L(\theta, a) = (\theta - a)^2$, y $\theta \sim N(\mu_0, 1)$.

(a) El estimador de Bayes posterior es la media de la distribución posterior de θ.

$$f(\theta, y) = \frac{(\frac{1}{\sqrt{2\pi}})^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \theta)^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu_0)^2\right)}{\int_{-\infty}^{\infty} (\frac{1}{\sqrt{2\pi}})^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \theta)^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu_0)^2\right) d\theta}.$$

Considerando $y_0 = \mu_0$:

$$f(\theta, y) = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2\right)}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2\right) d\theta},$$

$$f(\theta, y) = \frac{1}{\sqrt{\frac{2\pi}{n+1}}} \exp\left(-\frac{n+1}{2}\left(\theta - \sum_{i=1}^{n}\frac{y_i}{n+1}\right)^2\right),$$

$$E(\theta|y_1, y_2, ..., y_n) = \sum_{1=0}^{n}\frac{y_i}{n+1},$$

$$Var(\theta|y_1, y_2, ..., y_n) = \frac{1}{n+1}.$$

(b) Aproximación Bayesiana

Cuando $L(\theta, a) = (\theta - a)^2$, la regla de Bayes (o estimador de Bayes) es la media de $\prod(\theta|y) = P(\theta|y)$.

Por lo tanto; el estimador de Bayes o regla de Bayes con respecto a la pérdida cuadrada del error es:

$$\frac{y_0 + \sum_{i=1}^{n} y_i}{n+1} = \frac{\mu_0 + \sum_{i=1}^{n} y_i}{n+1}.$$

Es decir, en este caso, la decisión óptima que mínimiza la pérdida esperada es $\tilde{\theta} = E(\theta)$.

La mejor estimación de θ con pérdida cuadrática es la media de la distribución de θ en el momento de producirse la estimación.

Si $L(\theta, a) = w(\theta)(\theta - a)^2$, la regla de Bayes es:

$$d(y) = \frac{E^{\prod(\theta|y)}[\theta w(\theta)]}{E^{\prod(\theta|y)}[W(\theta)]},$$

$$d(y) = \frac{\int \theta w(\theta) f(y|\theta) d\theta}{\int w(\theta) f(y|\theta) d\theta}.$$

Si $L(\theta, a) = |q - a|^2$, cualquier mediana de $\prod(\theta|y)$ es un estimador de Bayes de θ.

Si $L(\theta, a) = \begin{cases} k_0(\theta - a), & si\ \theta - a \geq 0, \\ k_1(a - \theta), & si\ \theta - a < 0, \end{cases}$, cualquier $\frac{k_0}{K_0 + K_1}$ cuantil de $\prod(\theta|y)$ es un estimador de Bayes de θ.

4.3. Principios fundamentales de inferencia

4.3.1. El principio de verosimilitud

Para hacer inferencia sobre θ, después de haber observado Y, toda la información pertinente está contenida en la función de verosimilitud $l(\theta|y)$. Además, dos funciones de verosimilitud tienen la misma información sobre θ si son proporcionales entre sí (ver [18]).

Los métodos bayesianos cumplen el principio de verosimilitud: Si $l_1(\theta|y) \propto l_2(\theta|y)$ entonces, dada una distribución a priori $\pi(\theta)$,

$$f_1(\theta|y) \propto \pi(\theta) l_1(\theta|y) \propto \pi(\theta) l_2(\theta|y) \propto f_2(\theta|y).$$

4.3.2. Principio de suficiencia

Supongamos $Y = (y_1, ..., y_n)$ observaciones, que ayudan a ganar conocimiento acerca del parámetro θ, y que

$$t = t(y),$$

es una función de las observaciones. Llamaremos a esta función un estadístico. A menudo suponemos que t es un valor real, pero algunas veces tiene valores vectoriales.

$$p(y|\theta) = p(y, t|\theta) = p(t, \theta) p(y|t, \theta).$$

Sin embargo, algunas veces pasa que

$$p(y|t,\theta),$$

no depende de θ, es decir

$$p(y|\theta) = p(t|\theta)p(y|t).$$

Si esto pasa, decimos que t es un estadístico suficiente para θ dado y, algunas veces, abreviamos para decir que t es suficiente para θ.

Un estadístico $u = u(y)$ con densidad $p(u|\theta) = p(u)$ donde no depende de θ, es llamada un estadístico auxiliar para θ.

Teorema 4.1. *Un estadístico t es suficiente para θ dado y si y sólo si $l(\theta|y) \propto l(\theta|t)$ siempre que $t = t(y)$ (donde la constante de proporcionalidad no depende de θ).*

Demostración:

$\Rightarrow$) Si t es suficiente para θ dado y entonces

$$l(\theta|y) \propto p(y|\theta) = p(t|\theta)p(y|t) \propto p(t|\theta) \propto l(\theta|t).$$

$\Leftarrow$) Si tomamos la condición, entonces

$$p(y|\theta) \propto l(\theta|y) \propto l(\theta|t) \propto p(t|\theta)$$

tal que para alguna función $g(y)$

$$p(y|\theta) = p(t|\theta)g(y). \ \blacksquare$$

Corolario 4.1. *Para cada distribución a priori la distribución posterior de θ dada y es la misma como la distribución de θ dado un estadístico suficiente t.*

Demostración: Del teorema de Bayes $p(\theta|y)$ es proporcional a $p(\theta|t)$; entonce deben ser iguales a uno al integrar o sumar. $\blacksquare$

Corolario 4.2. *Si un estadístico $t = t(y)$ es tal que $l(\theta|y) \propto l(\theta|y')$ siempre que $t(y) = t(y')$, entonces t es suficiente para θ dado y.*

Demostración: Por sumar o integrar sobre todo y tal que $t(y) = t$,

$$l(\theta|t) \propto p(t|\theta) = \sum p(y'|\theta) \propto \sum l(\theta|y') \propto l(\theta|y).$$

La existencia de la suma sobre todo y' tal que $t(y') = t(y)$. ■

4.3.3. Teorema de factorización de Neyman

El siguiente teorema es usado frecuentemente para encontrar estadísticos suficientes.

Teorema 4.2. *Un estadístico t es suficiente para θ dado y si y sólo si existen funciones f y g tales que*

$$p(y|\theta) = f(t, \theta)g(y),$$

donde $t = t(y)$.

Demostración:

$\Rightarrow$) Si t es suficiente para θ dado y. Tenemos que

$$f(t, \theta) = p(t|\theta) \text{ y } g(y) = p(y|t).$$

$\Leftarrow$) Si tomamos la condición, integramos o sumamos ambos lados de la ecuación sobre todo los valores de y talque $t(y) = t$. Entonces el semiplano de la propiedad básica de densidad de t como el valor particular y tal que

$$p(t|\theta) = f(t, \theta)G(t).$$

Donde $G(t)$ es obtenida por sumar o integrar $g(y)$, sobre todos los valores de y. Obtenemos

$$f(t, \theta) = \frac{p(t|\theta)}{G(t)}.$$

Considerando ahora uno de los valores de y tal que $t(y) = t$ y sustituyendo en la ecuación declarada en el teorema, obtenemos

$$p(y|\theta) = p(t|\theta)\frac{g(y)}{G(t)}.$$

Por lo tanto, sea t suficiente o no,

$$p(y|t,\theta) = \frac{p(y,t|\theta)}{p(t|\theta)} = \frac{p(y|\theta)}{p(t\theta)},$$

vemos que

$$p(y|t,\theta) = \frac{g(y)}{G(t)}.$$

Por lo tanto, el semiplano derecho no depende de θ. De esto seguimos que t es ciertamente suficiente, y el teorema es probado. ■

4.3.4. El principio de condicionalidad

Suponiendo que se pueden hacer dos posibles experimentos E_1 y E_2 en relación a θ, y se elige uno con probabilidad 0.5, entonces la inferencia sobre θ depende sólo del resultado del experimento seleccionado.

El principio de verosimilitud equivale al principio de suficiencia más el principio de condicionalidad (ver [18]).

4.3.5. La familia exponencial

Se tienen distribuciones con estadísticas comunes que tienen una forma similar, la función de densidad de los parámetros pertenece a la familia exponencial si se puede escribir de tal forma

$$p(y|\theta) = g(y)h(\theta)\exp\{t(y)\psi(\theta)\},$$

o equivalentemente si la máxima verosimilitud de n observaciones independientes $y = (y_1, y_2, ..., y_n)$ de esta distribución es

$$l(\theta|y) \propto h(\theta)^n \exp\{\sum t(y_i)\psi(\theta)\}.$$

Si aplicamos inmediatamente la factorización de Neyman $\sum t(y_i)$, se observa que es suficiente para θ dado y.

4.3.6. La familia exponencial para dos parámetros

Los parámetros de la familia exponencial, como el nombre implica, sólo incluye funciones de densidad con parámetros conocidos.

Hay pocos casos en los que tenemos dos parámetros desconocidos, los más notorios son cuando la media y varianza de una distribución normal son ambos desconocidos, una función de densidad es de la forma de dos parámetros de la familia exponencial si es de la siguiente forma,

$$p(y|\theta,\phi) = g(x)h(\theta,\phi)\exp\left\{t(y)\psi(\theta,\phi) + u(y)\chi(\theta,\phi)\right\},$$

o equivalentemente si la máxima verosimilitud de n observaciones independientes $y = (y_1, y_2, ..., y_n)$ toma la forma

$$l(\theta,\phi|y) \propto h(\theta,\phi)^n \exp\left\{\sum t(y_i)\psi(\theta,\phi) + \sum u(y_i)\chi(\theta,\phi)\right\}.$$

Evidentemente el vector de dimensión dos $(\sum t(y_i), \sum u(y_i))$ es suficiente para el vector de dimensión dos (θ,ϕ) de parámetros dados. La familia de densidades conjugadas tal que su máxima verosimilitud toma la forma

$$p(\theta,\phi) \propto h(\theta,\phi)\exp\{\tau\psi(\theta,\phi) + \upsilon\chi(\theta,\phi)\}.$$

Mientras el caso de la distribución normal con ambos parámetros desconocidos es de considerable teoría y práctica importante. La idea de la familia exponencial puede ser extendida a k-parámetros (ver [18]).

4.4. Intervalo de credibilidad o regiones veraces

La idea de una región veraz o intervalo de credibilidad es proporcionar el análogo de un intervalo de confianza en estadística clásica. El razonamiento es que los estimados puntuales no proporcionan una medida de la precisión de la estimación ([32]). Esto causa problemas en la estadística clásica desde que los parámetros no son considerados como aleatorios, por lo tanto, no es posible dar un intervalo con la interpretación que existe una cierta probabilidad que el parámetro este en el intervalo.

En la teoría bayesiana, no hay dificultad para realizar esta aproximación porque los parámetros son tratados como aleatorios ([3]).

Definición 4.4. *Un conjunto veraz* $100(1-\alpha)\,\%$ *para* θ *es un subconjunto* C *de* Θ *tal que:*

$$1-\alpha \leq P(C|y) = \begin{cases} \int_C f(\theta|y)d\theta, & (caso\ continuo), \\ \sum_{\theta \epsilon C} f(\theta|y), & (caso\ discreto). \end{cases}$$

Un aspecto importante con los conjuntos veraces (y lo mismo sucede con los intervalos de confianza) es que ellos no son únicos.

Ya que la distribución a posteriori es actualmente una distribución de probabilidad en Θ, se puede decir subjetivamente que la probabilidad de θ está en C. Esta es la diferencia con los procedimientos de la estadística clásica los cuales sólo se interpretan como la probabilidad de que la variable aleatoria Y sea tal que, el conjunto creible $C(Y)$ contenga a θ.

Cualquier región con probabilidad $(1-\alpha)$ cumple la definición. Pero solamente se desea el intervalo que contiene únicamente los valores "más posibles" del parámetro, por lo tanto, es usual imponer una restricción adicional que indica, que el ancho del intervalo debe ser tan pequeño como sea posible ([3]).

Para hacer esto, uno debe considerar sólo aquellos puntos con las $f(\theta|y)$ más grandes. Esto conduce a un intervalo (o región) de la forma:

$$C = C_\alpha(y) = \{\theta : f(\theta|y) \geq \gamma\},$$

donde γ es elegido para asegurar que $\int_C f(\theta|y)d\theta = 1-\alpha$.

La región C que cumple las anteriores condiciones se denomina "región de densidad posterior más grande" (HPD), máxima densidad.

Generalmente, un HPD es encontrado por métodos numéricos, aunque para muchas distribuciones univariadas a posteriori, los valores de la variable aleatoria correspondientes son tabulados para un rango de valores de α.

Ejemplo:(media de una normal). Sean $\{y_1, ..., y_n\}$ **una muestra aleatoria de una distribución normal,**

$N(\theta, \sigma^2)$, con σ^2 conocido, con una a priori para θ de la forma: $\theta \sim N(b, d^2)$.

Se sabe que:

$$\theta|y \sim N\left(\frac{\frac{b}{d^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}\ ,\ \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}\right).$$

Lo que se quiere decir es que,

$$\left(\frac{\frac{b}{d^2}+\frac{n\bar{y}}{\sigma^2}}{\frac{1}{d^2}+\frac{n}{\sigma^2}}\right) \pm \mathcal{Z}_{\frac{\alpha}{2}} \left(\frac{1}{\frac{1}{d^2}+\frac{n}{\sigma^2}}\right)^{\frac{1}{2}}.$$

Cuando tomamos un n demasiado grande ($n \to \infty$), entonces tenenos que nuestro intervalo de credibilidad o veraz toma la siguiente forma ($\bar{y} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$).

Entonces el conjunto creible es igual al de estadística clásica. Pero sus interpretaciones son distintas, como se ve a continuación.

4.4.1. Diferencia entre los conjuntos creíbles y los intervalos de confianza

Bayesianamente un conjunto creíble C al $(1-\alpha)100\,\%$ significa que, la *probabilidad de que* θ *esté en* C *dados los datos observados es al menos* $(1-\alpha)100\,\%$. En otras palabras, un conjunto creíble da la certeza de que dada la muestra aleatoria tal conjunto C contenga a θ con al menos una probabilidad de $(1-\alpha)100\,\%$.

Clásicamente un intervalo de confianza al $(1-\alpha)100\,\%$ significa que, *si se pudiera calcular* C *para una gran cantidad de conjuntos de muestras aleatorias, cerca del* $(1-\alpha)100\,\%$ *de ellos cubrirían el valor real de* θ. Es decir, si se generan 100 conjuntos de muetras aleatorias, y se calcula el intervalo de confianza para cada caso (no necesariamente iguales para cada caso), al menos $(1-\alpha)100\,\%$ cubrirían a θ, sin embargo, la desventaja principal que tienen los intervalos de confianza clásicos es que no siempre se pueden generar tantas muestras aleatorias.

4.4.2. ¿Cómo se obtiene el intervalo de mínima longitud (máxima densidad)?

1. Localizar la moda de la función de densidad (posterior) de θ;

2. A partir de la moda, trazar líneas rectas horizontales en forma descendiente hasta que se acumule $(1-\alpha)$ de probabilidad.

Para describir el contenido inferencial de la distribución posterior de la cantidad de interés $p(\theta|D)$ (donde D es una conjunto de datos), es frecuentemente, conveniente dar regiones $R \subset \Theta$ de probabilidad dadas en virtud de $P(\theta|D)$ ([4]). Por ejemplo, la identificación de regiones que contienen el $50\,\%$, $90\,\%$, $95\,\%$, ó $99\,\%$ de probabilidad. Una región $R \subset \Theta$ tal que $\int_R p(\theta|D) = q$ (dado D, el valor verdadero de θ pertenezca a R con probabilidad q), se dirá que es una región q-créible 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.

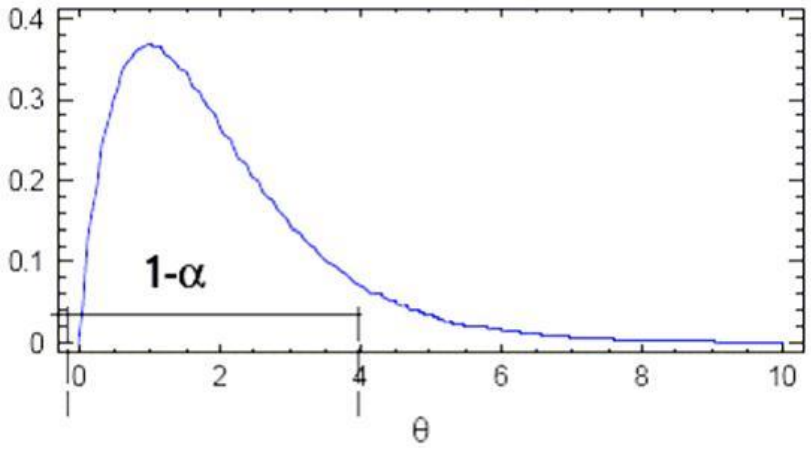

Figura 4.1: Distribución Gamma.

Claramente, para alguna q dada hay generalmente una infinidad de regiones de credibilidad pero, algunas veces, las regiones de credibilidad son seleccionadas para tener un tamaño mínimo, dando por resultado una región con función de densidad de probabilidad alta (HPD), donde todos los puntos en la región tienen una función de 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, $\theta_q = \theta_q(D)$ es el cuantil posterior $100_q\,\%$ de θ, $R = \{\theta; \theta \leq \theta_q\}$ es una región q-creíble unilateral, generalmente única. Más aún, las regiónes q-creíbles de probabilidad centrada de la forma $R = \left\{\theta; \theta_{\frac{1-q}{2}} \leq \theta \leq \theta_{\frac{1+q}{2}}\right\}$ son fácilmente calculables y son frecuentemente más empleados que las regiones HPD.

Análogamente a la estimación puntual, una región de credibilidad bayesiana se define como (ver [4]).

Definición 4.5. *Una región $R \subseteq \Theta$ tal que,*

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

se dirá 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 subconjuntos de $R \subseteq \Theta$ tal que $\int_R p(\theta)d\theta = 1 - \alpha$, para algún 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 el $100(1-\alpha)$ región R.

Proposición 4.1. *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 $A = \{R : P(\theta \in R) = 1 - \alpha\} \neq \emptyset$ y*

$L(R,\theta) = k\|R\| - I_{R(\theta)}$ con $R \in A$, $\theta \in \Theta$ y $k > 0$, entonces R es óptima 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 posiblemente en un subconjunto de Θ de probabilidad cero.

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

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

Definición 4.6. *Una función $R \subseteq \Theta$ será llamada una $100(1-\alpha)\,\%$ región con función de densidad de probabilidad alta para θ con respecto a $p(\theta)$ si:*

1. $P(\theta \epsilon R) = 1 - \alpha$.

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.*

4.5. Pruebas de hipótesis

Una hipótesis estadística consiste en una afirmación acerca de un parámetro. Lo que se desea saber al formular la hipótesis es saber si es verdadera o falsa. Usualmente, se pretende averiguar si es verdadera o falsa, se denomina hipótesis nula, esta se contrasta con la afirmación contraria a ella llamada hipótesis alternativa ([3]). En la estadística clásica para el contraste de hipótesis se tienen, la hipótesis *nula* $H_0 : \theta \in \Theta_0$ y la hipótesis alternativa $H_1 : \theta \in \Theta_1$. El procedimiento de contraste se realiza en términos de las probabilidades de *error tipo I* y *error tipo II*, estas probabilidades de error representan la posibilidad de rechazar la hipótesis nula dado que es verdadera o de no rechazar la hipótesis nula dado que es falsa, respectivamente ([18]).

En la Bayesiana, la tarea de decidir entre H_0 y H_1, es más simple. Pués solamente se calculan las probabilidades a posteriori $\alpha_0 = P(\Theta_0|y)$ y $\alpha_1 = P(\Theta_1|y)$ y se decide entre H_0 y H_1. La ventaja es que α_0 y α_1 son las probabilidades reales de las hipótesis de acuerdo a los datos observados y las opiniones a priori.

Se denotará con π_0 y π_1 a las probabilidades a priori de Θ_0 y θ_1 respectivamente.

Definición 4.7. *Se llamará razón a priori de H_0 contra H_1 al cociente $\frac{\pi_0}{\pi_1}$. Análogamente, razón a posteriori de H_0 contra H_1 será $\frac{\alpha_0}{\alpha_1}$.*

Si la razón a priori es cercana a la unidad significa que se considera que H_0 y H_1 tienen inicialmente (casi) la misma probabilidad de ocurrir. Si este cociente de probabilidades es grande significa que se considera que H_0 es más probable que H_1 y si el cociente es pequeño (< 1) es que inicialmente se considera que H_0 es menos probable que H_1.

Definición 4.8. *La cantidad*

$$B = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0},$$

es llamada ***factor de Bayes*** *en favor de* Θ_0.

En el interés del factor de Bayes, es que en algunas veces puede ser interpretado como la razón para H_0 con respecto a H_1, dado los datos. Esta interpretación es válida cuando de tienen hipótesis simples, esto es, $\Theta_0 = \{\theta_0\}$ y $\Theta_1 = \{\theta_1\}$, así se tiene

$$\alpha_0 = \frac{\pi_0 f(y|\theta_0)}{\pi_0 f(y|\theta_0) + \pi_1 f(y|\theta_1)},\ \alpha_1 = \frac{\pi_1 f(y|\theta_1)}{\pi_0 f(y|\theta_0) + \pi_1 f(y|\theta_1)},$$

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi f(y|\theta_0)}{\pi_1 f(y|\theta_1)},$$

$$B = \frac{\alpha_0\pi_1}{\alpha_1\pi_0} = \frac{f(y|\theta_0)}{f(y|\theta_1)}.$$

En otras palabras, B es la razón de probabilidad de H_0 contra H_1.

Además, nótese que el factor de Bayes proporciona una medida de la forma en que los datos aumentan o disminuyen las razones de probabilidades de H_0 respecto de H_1. Así, si $B > 1$ quiere decir que H_0 es ahora relativamente más probable que H_1 y si $B < 1$ ha aumentado la probabilidad relativa de H_1.

La distribución posterior $p(\theta|D)$, de la cantidad de interés θ, transmite de la 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 restricción $\theta \in \Theta_0 \subset \Theta$, 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 restringiendo Θ a Θ_0 simplificaría en gran medida el modelo, o porque existen además otros argumentos específicos en el contexto que sugieren que $\theta \in \Theta_0$. 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 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, de no rechazar (a_0) o rechazar (a_1) la restricción propuesta ([3]).

Para resolver este problema, es necesario especificar una apropiada función de pérdida $L(a_i, \theta)$, la cual mida las consecuencias de no rechazar o rechazar H_0 como una función del verdadero valor de θ. Note que esto requiere la especificación de una alternativa a_1 de no rechazar H_0; ésto sólo será esperado, para una acción tomada no porque sea buena, sino, porque es la mejor de todas.

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

$$\int_\Theta L(a_0, \theta) p(\theta|D) d\theta \; > \; \int_\Theta L(a_1, \theta) p(\theta|D) d\theta,$$

es más grande que la pérdida posterior esperada de rechazar:

$$\int_\Theta L(a_1, \theta) p(\theta|D) d\theta,$$

esto es, sí (y sólo si se cumple):

$$\int_\Theta \left[L(a_0, \theta) - L(a_1, \theta) \right] p(\theta|D) d\theta = \int_\Theta \Delta L(\theta) p(\theta|D) d\theta > 0.$$

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 contenido 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 ([3]).

4.5.1. Contraste de una cola o unilateral

Esto sucede cuando $\Theta \subseteq \mathbf{R}$ y $\Theta_0 \bigcup \Theta_1 \subset H = H_1 \bigcup H_2$, es decir.

$$H_0 : \theta \leq \theta_0 \ \ vs \ \ H_1 : \theta > \theta_0,$$

o el otro caso

$$H_0 : \theta \geq \theta_0 \ \ vs \ \ H_1 : \theta < \theta_0.$$

Lo interesante de este tipo de contraste es que el uso del valor p en el enfoque clásico tiene una justificación bayesiana (ver [3, 18]).

Ejemplo:

Supóngase que $Y \sim N(\theta, \sigma^2)$ y θ tiene una a priori no informativa $\pi(\theta) = 1$, entonces

$$\pi(\theta|y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\theta - y)^2}{2\sigma^2}\right\}.$$

El contraste a realizar es el siguiente:

$$H_0 : \theta \leq \theta_0 \ \ vs \ \ H_1 : \theta > \theta_0,$$

luego,

$$\alpha_0 = P(\theta \leq \theta_0|y) = \Phi((\theta_0 - y)/\sigma).$$

Clásicamente el valor p es la probabilidad, cuando $\theta = \theta_0$. Aquí el valor p es,

$$p = P(Y \geq y) = 1 - \Phi((y - \theta_0)/\sigma),$$

ya que la distribución normal es simétrica, se tiene que el valor p es el mismo que α_0.

4.5.2. Contraste de hipótesis nula puntual

El contraste de hipótesis nula puntual se refiere al siguiente contraste

$$H_0 : \theta = \theta_0 \;\; vs \;\; H_1 : \theta \neq \theta_0.$$

Es interesante porque a diferencia del contraste clásico, contiene nuevas características pero principalmente porque las respuestas en el enfoque bayesiano difieren radicalmente de las que da el enfoque clásico (ver [2]).

Casi nunca se da el caso en que $\theta = \theta_0$ exactamente. Es más razonable tener a la hipótesis nula como $\theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b)$, en donde $b > 0$ es alguna constante elegida de tal forma que para cada $\theta \in \Theta_0$ puede ser considerada indistinguible de θ_0.

Si b llega a ser muy grande, ya no es recomendable utilizar el contraste por hipótesis nula puntual, porque no da buenos resultados. Dado que se debe hacer el contraste $H_0 : \theta \in (\theta_0 - b, \theta_0 + b)$, se necesita saber cuando es apropiado aproximar H_0 por $H_0 : \theta = \theta_0$. Desde la perspectiva del análisis bayesiano, la única respuesta a esta pregunta es, *la aproximación es razonable si las probabilidades a posteriori de H_0 son casi iguales en las dos situaciones.* Una condición fuerte cuando éste sea el caso es que la función de distribución conjunta de probabilidad (verosimilitud) observada sea aproximadamente constante en $(\theta_0 - b, \theta_0 + b)$.

Para llevar a cabo el contraste bayesiano para una hipótesis nula puntual $H_0 : \theta = \theta_0$, no se utiliza una densidad a priori continua, ya que en cualquiera de estos casos θ_0 dará una probabilidad a priori cero. Se debe utilizar una distribución inicial mixta que asigne probabilidad de π_0 al punto θ_0 y distribuya el resto, $1 - \pi_0$ en los valores $\theta \neq \theta_0$, es decir, la densidad $\pi_1 g_1(\theta)$, con una densidad propia g_1, en donde $\pi_1 = 1 - \pi_0$.

Si $Y_1, ..., Y_n$ es una muestra aleatoria simple de la variable aleatoria Y, la densidad marginal de $Y = (Y_1, ..., Y_n)$ es,

$$m(y) = \pi_0 f(y|\theta_0) + (1 - \pi_0) m_1(y),$$

donde $m_1(y) = \int_{\theta \neq \theta_0} f(y|\theta) g_1(\theta) d\theta$ es la densidad bajo H_1. Entonces la probabilidad a posteriori de $\theta = \theta_0$ es:

$$\alpha_0 = P(\theta_0|y) = \frac{f(y|\theta_0)\pi_0}{m(y)} = \frac{f(y|\theta_0)\pi_0}{f(y|\theta_0)\pi_0 + (1-\pi_0)m_1(y)} = \frac{f(y|\theta_0)\pi_0}{f(y|\theta_0)\pi_0 + \pi_1 m_1(y)},$$

luego,

$$\frac{\alpha_0}{\alpha_1} = \frac{P(\theta_0|y)}{1-\pi(\theta_0|y)} = \frac{P(\theta_0|y)}{\frac{f(y|\theta_0)\pi_0+\pi_1 m_1(y)-f(y|\theta_0)\pi_0}{f(y|\theta_0)\pi_0+\pi_1 m_1(y)}} = \frac{f(y|\theta_0)\pi_0}{\pi_1 m_1(y)}.$$

Entonces el factor de Bayes de H_0 contra H_1 es,

$$B = \frac{\alpha_0\pi_1}{\alpha_1\pi_0} = \frac{\pi_0 f(y|\theta)\pi_1}{\pi_1 m_1(y)\pi_0} = \frac{f(y|\theta_0)}{m_1(y)}.$$

4.6. Prueba de hipótesis sobre una proporción

4.6.1. Proporción igual a una constante

Se valora la hipótesis $H_0 : P = P_0$ frente a la hipótesis $H_1 : P \neq P_0$.

Saber el valor de P_0 que se quiere valorar. Luego hay que dar la distribución a priori que se supone para el parámetro P. Como se vió en la (sección. 3.1.8) de enfoque bayesiano, se operará con la distribución Beta(a,b), donde los parámetros a y b pueden introducirse directamente, o bien estimarse a partir de la media y la desviación estándar de la distribución Beta. En cualquiera de los dos casos, la media debe ser aproximadamente igual a P_0, cumpliendo la condición ([18]):

$$\frac{a}{a+b} \approx P_0 \;\; media \approx P_0,$$

lo cual es enteramente razonable, ya que si el usuario pensara que la a priori de la media del parámetro fuera diferente de P_0, no tendría sentido para él hacer la prueba de hipótesis. Si no se verifica esta condición, o más específicamente no se cumple que:

$$\left| p_0 - \frac{a}{a+b} \right| < 0.015.$$

Con la condición $a, b > 0$ y $\sqrt{p_0} > 1$; asimismo, la media debe ser un valor entre 0 y 1, la desviación estándar será positiva y acotada superiormente por un valor en específico.

El procedimiento exige que se establezca también una probabilidad a priori (q) de la validez de H_0. En ese punto hay que relacionar e y f, donde e es el número de éxitos y f el de fracasos, $n = e + f$ experiencias.

Entonces, el factor de Bayes a favor de $H_0(BF)$ es

$$\frac{P_0^e(1 - P_0^f)B(a, b)}{B(a + e, b + f)},$$

donde $B(a, b)$ es la función Beta evaluada en (a, b); $a, b > 0$ ó $a, b \in \mathbb{R}$, la cuál se define del modo siguiente:

$$B(a, b) = \int_0^1 u^{a-1}(1 - u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)},$$

donde $\Gamma(x) = \int_0^\infty u^{x-1} \exp^{-u} du$.

Calcula también el factor de Bayes en contra de $H_0(BC)$,

$$BC = \frac{1}{BF},$$

y la probabilidad a posteriori de la veracidad de H_0:

$$P(H_0|datos) = \frac{qBF}{qBF + (1 - q)}.$$

Debe tenerse en cuenta que si el usuario no quiere "tomar partido" a priori sobre la veracidad de H_0, puede optar por informar que $q = 0.5$, lo cual hace que la probabilidad a posteriori solo quede en función de BF:

$$P(H_0|datos) = \frac{BF}{BF + 1}.$$

Debe advertirse que en el Marco Bayesiano, no se trata de "elegir" entre una hipótesis y otra, de rechazar o no la hipótesis nula, sino de identificar en qué medida una es más razonable que la otra; tanto el BF como la probabilidad a posteriori sirven a ese fin, aunque esta última sea más expresiva para la mayor parte de los usuarios.

4.6.2. Proporción dentro de un intervalo

En este caso se valora la hipótesis $H_0 : P \in [P_1, P_2]$ frente a la hipótesis complementaria $H_1 : P < P_1$ ó $P > P_2$.

Además de P_1 y P_2, deben proveerse a los parámetros en la distribución de interés. Los valores de a y b pueden introducirse directamente, o bien estimarse a partir de la media y la desviación estándar de la distribución beta ([18]).

El procedimiento exige que se establezca también una probabilidad a priori (q) de la validez de H_0.

La probabilidad de la hipótesis,

$$P(H_0) = F(P_2) - F(P_1),$$

donde $F(P)$ es la función de distribución a posteriori evaluada en P (área bajo la curva de densidad a posteriori que queda a la izquierda de P), el Factor de Bayes a favor de $H_0(BF)$ es

$$BF = \frac{P(H_0)}{P(H_1)},$$

y la probabilidad a posteriori de la veracidad de H_0 es

$$PP = \frac{qBF}{qBF + (1 - q)}.$$

Debe tomarse en cuenta que si el usuario no quiere "tomar partido" a priori sobre la validez de H_0, puede optar por informar que $q = 0.5$, lo cual hace que la probabilidad a posteriori solo quede en función de BF:

$$PP = \frac{BF}{BF + 1},$$

lo cual, a su vez equivale a que $PP = P(H_0)$.

4.7. Prueba de hipótesis sobre una diferencia de proporciones

4.7.1. Igualdad de proporciones

Se valora la hipótesis $H_0 : P_1 = P_2$ contra la hipótesis $H_1 : P_1 \neq P_2$.

Debe comenzarse dando la probabilidad a priori (q) de la validez de H_0. Hay que definir la distribución $Beta(a_1, b_1)$ a priori en el supuesto de que valga la hipótesis H_0. Esto, como en los casos anteriormente explicados, se puede hacer dando directamente los valores de los parámetros a_1 y b_1, o especificando la media y la desviación estándar de la distribución beta ([18]).

Asimismo, hay que definir del mismo modo, la distribución $Beta(a_2, b_2)$ a priori de la segunda proporción en el supuesto de que no vale H_0. Es decir, expresar nuestra convicción acerca del posible valor de P_2, supuesto que no coincide con P_1. Finalmente, se informarán los valores e_1 y f_1, donde e_1 es el número de éxitos y f_1 el de fracasos, en donde $n_1 = e_1 + f_1$ experiencias llevadas en el primer caso, así como e_2 y f_2, donde e_2 es el número de éxitos y f_2 el de fracasos, en donde $n_2 = e_2 + f_2$ experiencias llevadas en el segundo caso.

El factor de Bayes a favor de H_0:

$$BF = \frac{B(a_1 + e_1 + e_2, b_1 + f_1 + f_2)B(a_2, b_2)}{B(a_1 + e_1, b_1 + f_1)B(a_2 + e_2, b_2 + f_2)},$$

donde $B(x, y)$ denota la función beta evaluada en (x, y).

Se obtienen, asimismo, el factor de Bayes en contra de H_0:

$$BC = \frac{1}{BF},$$

y la probabilidad a posteriori de la veracidad de H_0:

$$P(H_0|datos) = \frac{qBF}{qBF + (1 - q)}.$$

4.7.2. Diferencia de proporciones dentro de un intervalo

Se valora la hipótesis $H_0 : P_1 - P_2 \in [P_3, P_4]$ contra la hipótesis complementaria: $H_1 : P_1 - P_2 < P_3 P_1 - P_2 > P_4$.

Debe comenzarse especificando los valores de P_3 y P_4, y definiendo las dos distribuciones a priori, $Beta(a_1, b_1)$ y $Beta(a_2, b_2)$. Esto, como en casos previos, se puede hacer dando directamente los valores de los parámetros a y b, o especificando la media y la desviación estándar de la distribución beta. Finalmente, se informarán los valores e_1 y f_1, donde e_1 es el número de éxitos y

f_1 el de fracasos, en donde $n_1 = e_1 + f_1$ experiencias llevadas en el primer caso, así como e_2 y f_2, donde e_2 es el número de éxitos y f_2 el de fracasos, en donde $n_2 = e_2 + f_2$ experiencias llevadas en el segundo caso ([18, 4]).

También hay que indicar la probabilidad a priori (q) de la validez de la hipótesis H_0.

La probabilidad de la hipótesis:

$$P(H) = F(P_4) - F(P_3),$$

el factor de Bayes a favor de H_0

$$BF = \frac{P(H_0)}{P(H_1)},$$

y la probabilidad a posteriori de la veracidad de H_0

$$PP = \frac{qBF}{qBF + (1-q)}.$$

4.8. Prueba de hipótesis sobre una media

4.8.1. Media igual a una constante

Se trata de valorar la hipótesis $H_0 : \mu = \mu_0$ frente a la hipótesis $H_1 : \mu \neq \mu_0$.

Los datos de entrada son el valor de μ_0 y la desviación estándar poblacional σ que se presuponen, la probabilidad a priori de la validez de H_0 y la desviación estándar de la distribución a priori bajo la hipótesis H_1. Finalmente, han de comunicarse los datos empíricos: media muestral y tamaño muestral: $\overline{x}$ y n ([18]).

El factor de Bayes a favor de la hipótesis H_0,

$$BF = \frac{\frac{\sqrt{n}}{\sigma} \exp\left[-\frac{n}{2\sigma^2}(\overline{x} - m_0)^2\right]}{\left(\frac{\sigma^2}{n} + t^2\right)^{-\frac{1}{2}} \exp\left[-\frac{(\overline{x}-m_0)^2}{2\left(\frac{\sigma^2}{n}+t^2\right)}\right]}.$$

El factor de Bayes en contra de H_0:

$$BC = \frac{1}{BF},$$

y la probabilidad a posteriori de la veracidad de H_0

$$P(H_0|datos) = \frac{qBF}{qBF + (1-q)}.$$

4.8.2. Media dentro de un intervalo

Se valora la hipótesis $H_0 : \mu \in [\mu_1, \mu_2]$ frente a la hipótesis $H_1 : \mu < \mu_1$ ó $\mu > \mu_2$.

Este procedimiento asume una distribución a priori uniforme (es decir, a priori no informativa). Las entradas son los valores de μ_1 y μ_2, la probabilidad a priori (q) de la validez de H_0 y, finalmente, los resultados empíricos: media, desviación estándar y tamaño muestrales: $\overline{x}$, s y n ([18]).

Se procede en forma similar cuando se estima la media para el caso uniforme; en ese caso, la distribución a posteriori será una normal con media m_p y desviación estándar s_p.

Las salidas que se producen son:

- Probabilidad de la hipótesis: $P(H_0) = F(\mu_2) - F(\mu_1)$, donde $F(\mu)$ es la distribución evaluada en μ, es decir, el área bajo la curva de densidad a posteriori (normal con media y desviación estándar m_p y s_p) que queda a la izquierda de μ.

- Factor de Bayes a favor de $H_0 : BF = \frac{P(H_0)}{P(H_1)}$.

- Probabilidad a posteriori de la veracidad de $H_0 : PP = \frac{qBF}{qBF+(1-q)}$.

4.9. Valoración de una prueba hipótesis sobre una diferencia de medias

4.9.1. Prueba de hipótesis sobre una diferencia de medias en dos poblaciones

A continuación se tratará el caso de dos muestras (respectivamente de cada población), donde aplicaremos la prueba de hipótesis para ambas muestras. La forma general de hacerlo es generalizando el factor de Bayes para el caso de dos muestras ([18]).

Ahora se considera la situación de muestras independientes con distribución normal;

$$x_1, x_2, ..., x_n \sim N(\lambda, \phi) \text{ y } y_1, y_2, ..., y_n \sim N(\mu, \psi),$$

que son independientes, aunque realmente el valor de interés es la distribución a posteriori de,

$$\delta = \lambda - \mu.$$

Antes de continuar, se debería tomar precauciones contra una posible aplicación inadecuada del modelo. Si $m = n$ y se define: $w_i = x_i - y_i$ y entonces investigar los w como una muestra $w_1, w_2, ..., w_n \sim N(\delta, \omega)$, para algún ω. Esto se conoce como el método de comparaciones pareadas.

En el caso del problema de dos muestras, se pueden presentar tres casos:

1. Cuando ϕ y ψ son conocidos;

2. Cuando se sabe que $\phi = \psi$ pero se desconocen sus valores;

3. Cuando ϕ y ψ son desconocidos.

Por los demás, de acuerdo a la naturaleza del trabajo, restringiremos nuestro trabajo al caso (1) cuando ϕ y ψ son conocidos. La razón principal para discutir este caso, es que el problema de la prueba implica menos complejidades en el caso donde las varianzas son desconocidas.

Si λ y μ tienen como referencia unas a prioris independientes (constantes), $p(\lambda) = p(\mu) \propto 1$ entonces como hemos visto anteriormente, con a priori normal, la distribución a posteriori para λ será $N(\overline{x}, \phi/m)$ y de forma similar la distribución a posteriori para μ será $N(\overline{y}, \psi/n)$ que es independiente de λ. De lo cual deducimos (veáse [18]):

$$\delta = \lambda - \mu \sim N(\overline{x} - \overline{y}, \phi/m + \psi/n).$$

El método se generaliza para este caso cuando la información importante esta disponible. Cuando la distribución a priori para λ es $N(\lambda_0, \phi_0)$ entonces la distribución a posteriori es:

$$\lambda \sim N(\lambda_1, \phi_1),$$

donde: $\phi_1 = \left(\phi_0^{-1} + (\frac{\phi}{m})^{-1}\right)^{-1}$ y $\lambda_1 = \phi_1 \left(\frac{\lambda_0}{\phi_0} + \frac{\overline{x}}{\frac{\phi}{m}}\right)$.

De modo semejante, si la distribución a priori para μ es $N(\mu_0, \varphi_0)$, entonces la distribución a posteriori para μ es $N(\mu_1, \varphi_1)$, donde φ_1 y μ_1 están definidos de modo semejante, como sigue:

$$\delta = \lambda - \mu \sim N(\lambda_1 - \mu_1, \phi_1 + \psi_1),$$

y las inferencias se proceden igual que antes.

4.9.2. Diferencia de medias dentro de un intervalo

Se considera la diferencia δ entre dos medias y se valora la hipótesis $H_0 : \delta \in [\delta_1, \delta_2]$ frente a la hipótesis $H_1 : \delta < \delta_1$ ó $\delta > \delta_2$.

Los datos de entrada son:

- Media, desviación estándar y tamaño muestrales en el grupo 1: $\overline{x}_1$, s_1 y n_1.
- Media, desviación estándar y tamaño muestrales en el grupo 2: $\overline{x}_2$, s_2 y n_2.
- Probabilidad a priori de la validez de $H_0 : q$.
- Probabilidad de la hipótesis, $P(H_0)$: se calcula la estimación no paramétrica de esta probabilidad a partir de la distribución empírica de las diferencias entre las dos medias.
- Factor de Bayes a favor de $H_0 : BF = \frac{P(H_0)}{P(H_1)}$.
- Probabilidad a posteriori de la veracidad de $H_0 : PP = \frac{qBF}{qBF+(1-q)}$.

4.10. Tablas de contingencia: prueba de hipótesis de independencia

La exploración de la relación entre mediciones categóricas es un problema básico de la inferencia.

La prueba estadística convencionalmente usada es la Ji-cuadrado de Pearson, y se rechaza la hipótesis de independencia si el valor del estadístico es suficientemente grande. Típicamente, se juzga el tamaño del valor del estadístico a través del valor p, probabilidad (bajo la hipótesis de independencia) de observar un valor del Ji-cuadrado con $(R-1) \times (C-1)$ grados de libertad al menos tan grande como el que se obtuvo, donde R es el número de filas de la tabla y C el de columnas. Si el valor de p es suficientemente pequeño, se rechaza la hipótesis de independencia.

La prueba bayesiana de independencia maneja dos hipótesis, la hipótesis H_0 que afirma que las dos variables son independientes, y la hipótesis complementaria, que plantea que las dos variables son dependientes o están de alguna manera relacionadas. Bajo el enfoque bayesiano, se valora el

apoyo relativo de los datos a cada una de estas dos hipótesis en términos del factor de Bayes, el cual compara las probabilidades de los datos observados bajo las hipótesis:

$$FB = \frac{P(D = d_0|H_0)}{P(D = d_0|H_1)}.$$

Naturalmente, mientras H_0 es una afirmación concreta , la falsedad de H_0 puede asumir infinitas expresiones puntuales. Sin embargo, en condiciones bastante generales, se pueden hallar cotas inferiores para BF (ver [4]). Así, para el caso en que se valora una asociación a través de la prueba Jicuadrado, se tiene una cota inferior para el factor de Bayes en función del valor observado χ^2:

$$\sqrt{\chi^2}\exp\left(\frac{1-\chi^2}{2}\right) \leq FB.$$

Esto da lugar a la siguiente desigualdad

$$\left[1 + \frac{1 - P(H_0)}{P(H_0)\sqrt{\sqrt{\chi^2}\exp\left(\frac{1-\chi^2}{2}\right)}}\right]^{-1} \leq P(H_0|D = d_0).$$

4.11. Inferencia predictiva

Los problemas predictivos en estadística son aquellos en los cuales las cantidades desconocidas de interés son las futuras variables aleatorias.

Una situación típica de este tipo es cuando la variable aleatoria Z, tiene una densidad $g(z|\theta)$ (θ desconocida), y se quiere predecir un valor de Z cuando existen datos, y, derivados de una densidad $h(y|\theta)$. Por ejemplo, y pueden ser los datos de un estudio de regresión, y se desea predecir una futura respuesta de la variable aleatoria Z, en la regresión ([4]).

La idea es predecir a la variable aleatoria $Z \sim g(z|\theta)$ basado en la observación de $Y \sim h(y|\theta)$.

Ahora, se supone que Y y Z son independientes y g es una densidad (si no fueran independientes, sería necesario cambiar $g(z|\theta)$ por $g(z|\theta, y)$). La idea de la predicción bayesiana es que, ya que $\pi(\theta|y)$ es la distribución a posteriori de θ, entonces $g(z|\theta)\pi(\theta|y)$ es la distribución conjunta de Z y θ dado Y, y al integrar sobre θ se obtiene la distribución aposteriori de Z dado Y.

Definición 4.9. *La **densidad predictiva** de Z dado $Y = y$, cuando la distribución a priori para θ es π, está definida por*

$$P(z|y) = \int_\theta g(z|\theta)\pi(\theta|y)d\theta.$$

Consideremos el siguiente modelo de regresión lineal.

$$Z = \theta Y + \varepsilon,$$

donde $\varepsilon \sim N(0, \sigma^2)$ con σ^2 conocido, y supóngase que se tiene un conjunto de datos disponibles $((z_1, y_1), ..., (z_n, y_n))$.

Un estadístico suficiente para θ, es el estimador por mínimos cuadrados X, en donde

$$X = \frac{\sum_{i=1}^{n} Z_i y_i}{\sum_{i=1}^{n} y_i^2}.$$

Como Z_i tiene una fdp normal y X es combinación lineal de normales, entonces X tendrá una distribución normal.

$$X \sim N(\theta, \sigma^2/S_y), \;\; en \;\; donde: \;\; S_y = \sum_{i=1}^{n} y_i^2.$$

Si se utiliza una a priori no informativa $\pi(\theta) = 1$ para θ, la distribución a posteriori, $\pi(\theta|x) \sim N(x, \sigma^2/S_y)$.

Ahora se obtiene la probabilidad de Z dado X:

$$\begin{aligned} P(z|x) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(z-\theta y)^2\right\} \frac{1}{\sqrt{2\pi}\sigma/\sqrt{S_y}} \times \\ &\qquad \exp\left\{-\frac{1}{2\sigma^2/S_y}(\theta - x)^2\right\} d\theta \\ &= \frac{\sqrt{S_y}}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}[(z-\theta y)^2 + S_y(\theta - x)]^2\right\} d\theta. \end{aligned}$$

Desarrollando separadamente $[(z-\theta y)^2 + S_y(\theta - x)]^2$, se tiene:

$$\begin{aligned} [(z-\theta y)^2 + S_y(\theta - x)]^2 &= z^2 + S_y x^2 + \theta^2 y^2 + S_y \theta^2 - 2S_y x\theta + S_y x^2 \\ &= z^2 + S_y x^2 + \theta(y^2 + S_y) - 2\theta(yz + S_y x) \\ &= (y^2 + S_y)\left(\frac{z^2 + S_y x^2}{y^2 + S_y} - \frac{(yz + S_y x)^2}{(y^2 + S_y)^2} + \left(\theta - \frac{yz + S_y}{y^2 + S_y}\right)^2\right). \end{aligned}$$

Luego sustituyendo nuevamente la expresión anterior en $P(z|x)$,

$$P(z|x) = \frac{\sqrt{S_y}}{2\pi\sigma^2} \exp\left\{-\frac{(y^2+S_y)}{2\sigma^2}\left[\frac{z^2+S_yx^2}{y^2+S_y} - \frac{(yz+S_yx)^2}{(y^2+S_y)^2}\right]\right\} \times$$

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2/(y^2+S_y)}\left(\theta - \frac{yz+S_y}{y^2+S_y}\right)^2\right\} d\theta$$

$$= \frac{\sqrt{S_y}}{2\pi\sigma^2} \exp\left\{-\frac{(y^2+S_y)}{2\sigma^2}\left[\frac{z^2+S_yx^2}{y^2+S_y} - \frac{(yz+S_yx)^2}{(y^2+S_y)^2}\right]\right\} \left(\frac{\sigma}{\sqrt{y^2+S_y}}2\pi\right)$$

$$= \frac{1}{\sqrt{2\pi}\sigma\sqrt{\frac{y^2+S_y}{S_y}}} \exp\left\{\frac{1}{2\sigma^2\left(\frac{y^2+S_y}{S_y}\right)}(z-xy)^2\right\}.$$

Finalmente se tiene,

$$P(z|x) = N\left(xy, \frac{\sigma^2(y^2+S_y)}{S_y}\right).$$

Capítulo 5

Regresión Lineal Bayesiano

El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables. Las aplicaciones de la regresión se encuentran numerosos campos de la investigación como en ingeniería, ciencias físicas, químicas, biológicas etc. De hecho es una de las técnicas estadísticas más utilizadas.

Los modelos de regresión son empleados para:

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

5.1. Regresión lineal simple

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

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{5.1}$$

donde β_0 la constante interceptora y β_1 la pendiente, son constantes desconocidas, y ϵ es un error aleatorio. La componente aleatoria se asume que tiene media cero y varianza constante desconocida, adicionalmente que no están 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:

$$E[y|x] = \beta_0 + \beta_1 x, \tag{5.2}$$

y la varianza,

$$V[y|x] = V[\beta_0 + \beta_1 + \epsilon] = \sigma^2. \tag{5.3}$$

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 de 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. Uno de los métodos más usados para estimar dichos valores es el de **mínimos cuadrados** que no es más que la suma mínima de los cuadrados de las diferencias entre observaciones y_i y la línea recta. A saber, los estimadores por mínimos cuadrados de la intercepción y la pendiente son:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{5.4}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}, \tag{5.5}$$

donde $\bar{y}$ y $\bar{x}$ 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.

5.1.1. Estimación mediante la función de verosimilitud

Sean y_i y x_i los datos observados, para $i = 1, 2, ..., n$. Por simplicidad matemática, no se toma en cuenta la intersección, el modelo de regresión lineal es:

$$y_i = \beta x_i + \epsilon_i. \tag{5.6}$$

Los supuestos acerca de ϵ_i y x_i determinan la forma de la función de verosimilitud, estos son:

1. ϵ_i se distribuye en forma de una normal con media 0 y varianza σ^2, ϵ_i y ϵ_j son independientes el uno del otro para $i \neq j$. Esto es, ϵ_i es independiente e idénticamente distribuido (i.i.d.) a $N(0, \sigma^2)$.

2. x_i es fijo, y si son variables aleatorias, son independientes de ϵ_i con una función de densidad de probabilidad, $p(x_i|\lambda)$ donde λ es un vector de parámetros que no incluyen a β y σ^2.

La suposición de que las variables explicativas no son aleatorias es común en las ciencias físicas, donde los métodos experimentales son comunes. Es decir, como parte de la disposición experimental, el investigador escoge valores particulares para las x. En muchas aplicaciones económicas, tal suposición no es razonable. Sin embargo, la suposición de que la distribución de las x es independiente del error y con una distribución, que no depende de los parámetros de interés, es a menudo razonable.

La función de verosimilitud está definida como la función de densidad de probabilidad conjunta para todos los datos en los parámetros desconocidos. El vector de observaciones de la variable dependiente es un vector de longitud n:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix},$$

o, equivalentemente $\mathbf{y} = (y_1, y_2, ..., y_n)'$. De manera similar, para la variable explicativa, definimos $\mathbf{x} = (x_1, x_2, ..., x_n)'$. Entonces la función de verosimilitud llega a ser $p(\mathbf{y}, \mathbf{x}|\beta, \sigma^2, \lambda)$. El segundo supuesto implica que podemos escribir la función de verosimilitud como:

$$p(\mathbf{y}, \mathbf{x}|\beta, \sigma^2, \lambda) = p(\mathbf{y}|\mathbf{x}, \beta, \sigma^2)p(\mathbf{x}|\lambda).$$

Como la distribución de x no es de interés, se trabaja entonces con una función de verosimilitud sobre $p(\mathbf{y}|\mathbf{x}, \beta, \sigma^2)$.

Los supuestos sobre los errores pueden ser usados para trabajar en la forma precisa de la función de verosimilitud. En particular, usando ciertas reglas básicas de probabilidad , encontramos:

- $p(y_i|\beta, \sigma^2)$ es normal.
- $E(y_i|\beta, \sigma^2) = \beta x_i$.
- $Var(y_i|\beta, \sigma^2) = \sigma^2$.

Usando la definición de la densidad de una distribución normal obtenemos:

$$p(y_i|\beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right].$$

Finalmente, para $i \neq j$, si ϵ_i y ϵ_j son independientes, se sigue que y_i y y_j son también independientes y así, $p(\mathbf{y}|\beta, \sigma^2) = \prod_{i=1}^{n} p(y_i|\beta, \sigma^2)$, por lo tanto, la función de verosimilitud está dada por,

$$p(\mathbf{y}|\beta, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right]. \tag{5.7}$$

Si se toma

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \beta x_i)^2 &= \sum_{i=1}^{n}\left\{(y_i - \hat{\beta}x_i) - (\beta - \hat{\beta})x_i\right\}^2 \\
&= \sum_{i=1}^{n}\left\{(y_i - \hat{\beta}x_i)^2 + (\beta - \hat{\beta})^2 x_i^2\right\} \\
&= \sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2 + (\beta - \hat{\beta})^2\sum_{i=1}^{n}x_i^2 \\
&= (n-1)\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2}{(n-1)} + (\beta - \hat{\beta})^2\sum_{i=1}^{n}x_i^2 \\
&= vs^2 + (\beta - \hat{\beta})^2\sum_{i=1}^{n}x_i^2.
\end{aligned}$$

Entonces para futuras derivaciones, la verosimilitud se escribirá como (ver [15]),

$$\sum_{i=1}^{n}(y_i - \beta x_i)^2 = vs^2 + (\beta - \hat{\beta})^2\sum_{i=1}^{n}x_i^2,$$

donde,

$$v = (n-1), \;\; \hat{\beta} = \frac{\sum_{i=1}^{n}x_i y_i}{\sum_{i=1}^{n}x_i^2}, \;\; y \;\; s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2}{v},$$

además, $\hat{\beta}$, s^2 y v son los estimadores de mínimos cuadrados ordinarios (MCO) para β, el error estándar y los grados de libertad, respectivamente. Son estadísticas suficientes (ver [24]). Además, para muchas derivaciones técnicas, es más fácil trabajar con la precisión del error que con la varianza. La precisión de error está definida como $\tau = \frac{1}{\sigma^2}$.

Usando estos resultados, podemos escribir la función de verosimilitud como:

$$p(y|\beta, \tau) = \frac{1}{2\pi^{\frac{n}{2}}}\left(\tau^{\frac{1}{2}}\exp\left[-\frac{\tau}{2}(\beta - \hat{\beta})^2\sum_{i=1}^{n}x_i^2\right]\right)\left(\tau^{\frac{v}{2}}\exp\left[-\frac{\tau v}{2s^{-2}}\right]\right). \tag{5.8}$$

El primer término en los corchetes es el núcleo de la densidad normal para β, y el segundo término es una densidad gamma para τ.

5.1.2. Función a priori del modelo de regresión lineal simple

La *a priori* refleja cualquier información que el investigador tiene antes de ver los datos, que desea incluir. Por lo tanto, los previos pueden tomar cualquier forma. Sin embargo, es común escoger clases particulares de previos que son fáciles de interpretar y/o hacer los cálculos más fáciles. La *a priori* conjugada natural típicamente tiene tales ventajas. Una distribución *a priori* conjugada es una que, cuando es combinada con la verosimilitud, produce una posterior que tiende en la misma clase de distribución. Estas propiedades significan que la información *a priori* puede interpretarse del mismo modo que la función de verosimilitud.

En el modelo de regresión lineal simple, se extrae una *a priori* para β y τ, que se denota por $p(\beta, \tau)$. La densidad posterior se denotará por $p(\beta, \tau|\mathbf{y})$, es conveniente escribir $p(\beta, \tau) = p(\beta|\tau)p(\tau)$ y pensar en términos de un previo para $\beta|\tau$ y uno para τ. La forma de la función de verosimilitud en (4.11) sugiere que el previo conjugado natural incluirá una distribución *normal* para $\beta|\tau$ y una distribución *gamma* para τ. A la distribución final, es el producto de una *gamma* y una *normal*, se le denominara una $normal - gamma$, (veáse [15, 26]).

$$\beta|\tau \sim N(\beta, \tau^{-1}\underline{V}) \;\; y \;\; \tau \sim G(\underline{s}^{-2}, \underline{v}),$$

$$\beta, \tau \sim NG(\underline{\beta}, \underline{V}, \underline{s}^2, \underline{v}). \tag{5.9}$$

El investigador podrá escoger valores particulares de los llamados hiperparámetros previos $\underline{\beta}, \underline{V}, \underline{s}^2$ y $\underline{v}$ para reflejar información previa.

Para esta ocasión se usará la barra bajo el parámetro ($\underline{\beta}$) para denotar los parámetros de una densidad previa, y la barra sobre el parámetro ($\overline{\beta}$) para denotar los parámetros de una densidad posterior.

5.1.3. Función a posteriori del modelo de regresión lineal simple

La densidad posterior resume la información *a priori* de los datos, que tenemos sobre los parámetros desconocidos, β y τ. La densidad posterior es también de la forma $normal - gamma$, confirmando que la *a priori* es en realidad una conjugada natural.

Formalmente, tenemos la posterior de la forma, (ver [15, 26])

$$\beta, \tau|y \sim NG(\overline{\beta}, \overline{V}, \overline{s}^{-2}, \overline{v}), \tag{5.10}$$

donde

$$\overline{V} = \frac{1}{\underline{V}^{-1} + \sum_{i=1}^{n} x_i^2}, \;\; \overline{\beta} = \overline{V}(\underline{V}^{-1}\underline{\beta} + \hat{\beta}\sum_{i=1}^{n} x_i^2), \;\; \overline{v} = \underline{v} + n,$$

y $\overline{s}^{-2}$ está definido a través de

$$\overline{vs}^2 = \underline{vs} + vs^2 + \frac{(\hat{\beta} - \underline{\beta})^2}{\underline{V} + \left(\frac{1}{\sum_{i=1}^{n} x_i^2}\right)}. \tag{5.11}$$

En el modelo de regresión, el coeficiente de la variable explicativa β, es una medida de los efectos marginales de la variable explicativa en la variable dependiente. La media posterior $E(\beta|\mathbf{y})$, es un punto de estimación y $Var(\beta|\mathbf{y})$ es usado para la medida de la incertidumbre asociada con el punto de estimación. Usando las reglas básicas de probabilidad, la media posterior puede ser calculada como:

$$E(\beta|\mathbf{y}) = \int\int \beta p(\beta, \tau|\mathbf{y}) \; d\tau \; d\beta = \int \beta p(\beta|\mathbf{y}) \; d\beta.$$

Esta ecuación motiva el interés sobre la densidad marginal posterior $p(\beta|\mathbf{y})$. Puede ser calculado analíticamente usando las propiedades de la distribución $Normal - Gamma$. En particular, implica que, si se integra con respecto a τ (usando el hecho de que $p(\beta|\mathbf{y}) = \int p(\beta, \tau|\mathbf{y})d\tau$), la distribución marginal posterior para β es una distribución t, (ver [15, 26]).

$$\beta|\mathbf{y} \sim t(\overline{\beta}, \overline{s}^2\overline{V}, \overline{v}), \tag{5.12}$$

se sigue de la definición de la distribución t, que la

$$E(\beta|\mathbf{y}) = \overline{\beta}, \tag{5.13}$$

y

$$Var(\beta|\mathbf{y}) = \frac{\overline{vs}^2}{\overline{v} - 2}\overline{V}. \tag{5.14}$$

La precisión del error τ, es usualmente de menos interés que β, pero las propiedades de la $normal - gamma$ implican inmediatamente que (ver [25]):

$$\tau|\mathbf{y} \sim G(\overline{s}^{-2}, \overline{v}), \tag{5.15}$$

y por lo tanto,

$$E(\tau|\mathbf{y}) = \overline{s}^{-2}, \tag{5.16}$$

y

$$Var(\tau|\mathbf{y}) = \frac{2\overline{s}^{-2}}{\overline{v}}. \tag{5.17}$$

El modelo de regresión lineal con la función inicial conjugada $normal - gamma$, es un caso donde la simulación posterior no es requerida.

Para ver las diferencias entre la Bayesiana y Clásica, tómese en cuenta que este último podría calcular $\hat{\beta}$ y su varianza $s^2/(\sum_{i=1}^{n} x_i^2)$, y estimar σ^2 por s^2. Los Bayesianos calculan la media y la varianza posterior de β por ($\overline{\beta}$ y $\frac{\overline{v}\overline{s}^2}{\overline{v}-2}\overline{V}$) y se estima $\tau = \sigma^{-2}$ por su media posterior $\overline{s}^{-2}$. Éstas son estrategias muy similares, si no fuera por dos diferencias importantes. En primer lugar, la fórmula Bayesiana combina la *a priori* y la información de los datos. En segundo término, está la interpretación Bayesiana de β como una variable aleatoria.

Tomando $\underline{v}$ relativamente pequeño, n y $\underline{V}$ con valores grandes que asegure que la información previa juegue un papel pequeño en la fórmula posterior (como (5.14)-(5.17)). Se refiere como una inicial relativamente no informativa.

Se establece una inicial no informativa tomando $\underline{v} = 0$ y $\underline{V} = 0$. Tales elecciones son hechas comúnmente, e implican que $\beta, \tau|\mathbf{y} \sim NG(\overline{\beta}, \overline{V}, \overline{s}^{-2}, \overline{v})$ (ver [15]), donde

$$\overline{V} = \frac{1}{\sum_{i=1}^{n} x_i^2}, \;\; \overline{\beta} = \hat{\beta}, \;\; \overline{v} = N, \;\; y \;\; \overline{v}\overline{s}^2 = vs^2.$$

Que son los resultados de mínimos cuadrados ordinarios (MCO).

Las funciones iniciales no informativas tienen propiedades muy atractivas y dada la relación cercana con los resultados de MCO, proporciona un puente entre los enfoques Bayesiano y Clásico. Sin embargo, tienen una propiedad indeseable: esta densidad inicial no es una densidad válida, pués hace que no integre a uno, tales iniciales son denominadas impropias.

5.2. Regresión lineal múltiple con apriori conjugada natural

Una discusión detallada del modelo de regresión puede encontrarse en cualquier libro de econometría [15, 26]. Se tiene una variable dependiente y_i y k variables explicativas, $x_{i1}, ..., x_{ik}$ para $i = 1, ..., n$. El modelo de regresión lineal esta dado por:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \;\; x_{i0} = 1 \;\; para \;\; i = 1, 2, ...n. \tag{5.18}$$

Se definen los siguientes vectores $n \times 1$:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix},$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix},$$

el vector $k \times 1$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix},$$

y la matriz $n \times k$ se escribe como:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & . & . & . & a_{1k} \\ 1 & x_{22} & . & . & . & x_{2k} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n2} & . & . & . & x_{nk} \end{bmatrix}$$

y se escribe

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{5.19}$$

5.2.1. Función de verosimilitud

Los supuestos sobre $\boldsymbol{\epsilon}$ y $\mathbf{X}$ determinan la forma de la función de verosimilitud. Las generalizaciones son:

1. $\boldsymbol{\epsilon}$ tiene una distribución normal multivariada con media 0_n, donde $0_n = (0, 0, 0, ..., 0)'$ es un vector de $n - ceros$ y matriz de covarianzas $\sigma^2\mathbf{I}_n$, donde $\mathbf{I}_n$ es la matriz identidad $n \times n$. Es decir, $\boldsymbol{\epsilon} \sim N(0_n, \tau^{-1}\mathbf{I}_n)$ donde $\tau = \sigma^{-2}$.

2. Todos los elementos de $\mathbf{X}$ son fijos y si son variables aleatorias, estos son independientes de todos los elementos de $\boldsymbol{\epsilon}$ con una función de densidad de probabilidad $p(\mathbf{X}|\boldsymbol{\lambda})$, donde $\boldsymbol{\lambda}$ es un vector de parámetros que no incluyen a β ni a τ.

La matriz de covarianzas es una matriz que contiene las varianzas en la diagonal y las covarianzas fuera de esta, esto significa:

$$var(\boldsymbol{\epsilon}) = \begin{bmatrix} var(\epsilon_1) & cov(\epsilon_1,\epsilon_2) & . & . & cov(\epsilon_1,\epsilon_n) \\ cov(\epsilon_1,\epsilon_2) & var(\epsilon_2) & . & . & cov(\epsilon_2,\epsilon_n) \\ . & . & . & . & . \\ . & . & . & . & . \\ cov(\epsilon_n,\epsilon_1) & . & . & . & var(\epsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \tau^{-1} & 0 & . & . & 0 \\ 0 & \tau^{-1} & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \tau^{-1} \end{bmatrix},$$

$var(\boldsymbol{\epsilon}) = \tau^{-1}\mathbf{I}_n$ ó $var(\epsilon_i) = \tau^{-1}$ y $cov(\epsilon_i,\epsilon_j) = 0$ para $i,\ j = 1,...,n$ y $i \neq j$.

Usando la definición de densidad normal multivariada, podemos escribir la función de verosimilitud como:

$$p(\mathbf{y}|\boldsymbol{\beta},\tau) = \frac{\tau^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}}\left\{\exp\left[-\frac{\tau}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]\right\}. \tag{5.20}$$

Es conveniente escribir la función de verosimilitud desde el punto de vista de las cantidades de MCO. Éstos son [15]:

$$\mathbf{v} = n - k, \tag{5.21}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{5.22}$$

y

$$s^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{v}. \tag{5.23}$$

La función de verosimilitud se escribe como

$$p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\tau}) = \frac{1}{(2\pi)^{\frac{n}{2}}}\left\{\tau^{\frac{1}{2}}\exp\left[-\frac{\tau}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\}\left\{\tau^{\frac{v}{2}}\left[-\frac{\tau v}{2s^2}\right]\right\}. \tag{5.24}$$

5.2.2. Función a priori con conjugada natural

La *a priori* para β condicional en τ tiene la forma [15, 26],

$$\boldsymbol{\beta}|\tau \sim N(\underline{\boldsymbol{\beta}}, \tau^{-1}\underline{\mathbf{V}}),$$

y para τ es de la forma

$$\tau \sim G(\underline{s}^{-2}, \underline{v}),$$

entonces la posterior tiene la forma

$$\boldsymbol{\beta}, \boldsymbol{\tau} \sim NG(\boldsymbol{\beta}, \underline{\mathbf{V}}, \underline{s}^{-2}, \underline{v}), \tag{5.25}$$

donde, $\underline{\boldsymbol{\beta}}$ es ahora un k vector que contiene las medias previas para los k coeficientes de regresión, $\beta_1, ..., \beta_k$ y $\underline{\mathbf{V}}$ es ahora una matriz de covarianzas positiva definida por $k \times k$. La notación para la densidad es $p(\beta|\tau) = f_{NG}(\beta, \tau|\underline{\boldsymbol{\beta}}, \underline{\mathbf{V}}, \underline{s}^{-2}, \underline{v})$.

5.2.3. Función posterior con inicial natural

Esta se deriva multiplicando la verosimilitud (5.24) por los previos (5.25) produciendo una posterior de la forma [15, 26],

$$\boldsymbol{\beta}, \tau|\mathbf{y} \sim NG(\overline{\boldsymbol{\beta}}, \overline{\mathbf{V}}, \overline{s}^{-2}, \overline{v}), \tag{5.26}$$

donde

$$\overline{\mathbf{V}} = (\underline{\mathbf{V}}^{-1} + \mathbf{X}'\mathbf{X})^{-1}, \;\; \overline{\boldsymbol{\beta}} = \overline{\mathbf{V}}(\underline{\mathbf{V}}^{-1}\underline{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}), \;\; \overline{v} = \underline{v} + n,$$

y s^{-2} esta definido completamente por

$$\overline{vs}^2 = \underline{vs}^2 + vs^2 + (\hat{\boldsymbol{\beta}} - \underline{\boldsymbol{\beta}})' \left[\underline{\mathbf{V}} + (\mathbf{X}'\mathbf{X})^{-1}\right]^{-1} (\hat{\boldsymbol{\beta}} - \underline{\boldsymbol{\beta}}). \tag{5.27}$$

Las expresiones anteriores describen la distribución posterior conjugada. En el caso de la posterior marginal para β, el resultado es una distribución t multivariada [15],

$$\boldsymbol{\beta}|\mathbf{y} \sim t(\underline{\boldsymbol{\beta}}, \overline{s}^2\overline{\mathbf{V}}, \overline{v}), \tag{5.28}$$

y se sigue de la definición de la distribución t que:

$$E(\boldsymbol{\beta}|\mathbf{y}) = \overline{\boldsymbol{\beta}}, \;\; var(\boldsymbol{\beta}|\mathbf{y}) = \frac{\overline{v}\overline{s}^2}{\overline{v}-2}\overline{\mathbf{V}}.$$

Las propiedades de la distribución normal-gamma implican inmediatamente esto [25],

$$\tau|y \sim G(\overline{s}^{-2}, \overline{v}), \tag{5.29}$$

y por lo tanto,

$$E(\tau|y) = \overline{s}^{-2}, \tag{5.30}$$

$$var(\tau|y) = \frac{2\overline{s}^{-2}}{\overline{v}}. \tag{5.31}$$

Para una a priori no informativa, se toma un valor más pequeño para $\underline{v}$ que n y $\underline{\mathbf{V}}$ un valor grande. Cuando se trabaja con matrices, la interpretación del término grande no es inmediatamente obvia. Se toman A y B donde $A > B$ y A, B son matrices cuadradas, $A - B$ es positiva definida. Una medida de la magnitud de una matriz es su determinante. Por lo tanto, cuando decimos que A debe ser relativamente más grande que B, quiere decir que $A - B$ debe ser una matriz positiva definida con un determinante grande.

Se puede crear una función apriori no informativa tomando $\underline{v}$ y $\underline{\mathbf{V}}$ un valor pequeño. No existe una vía única de hacer esto último. Una vía común es poner $\underline{\mathbf{V}}^{-1} = c\mathbf{I}_k$, donde la c es un escalar, e $\mathbf{I}_k$ es la matriz identidad de k valores, dejamos entonces que c tienda a cero. Si se hace esto se encuentra [15, 26],

$$\overline{\mathbf{V}} = (\mathbf{X}'\mathbf{X})^{-1}, \;\; \overline{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}, \;\; \overline{v} = n \;\; y \;\; \overline{v}\overline{s}^2 = vs^2.$$

Todas estas fórmulas suponen información de los datos, y son iguales a las cantidades de mínimos cuadrados ordinarios.

En cuanto el caso de una variable explicativa, esta *a priori* no informativo es impropio y puede ser escrito como [26]:

$$p(\beta|\tau) \propto \frac{1}{\tau}. \tag{5.32}$$

Capítulo 6

Contaminación de Ríos

Se piensa que la contaminación del agua es causada más por actividades humanas que cualquier otra razón, ya que existen dos principales tipos de razones; las fuentes puntuales y las difusas [1]. Las primeras se encargan de implantar substancias contaminantes en localizaciones específicas a través de tuberías, por ejemplo, de alcantarillas en el agua superficial.

El segundo tipo de fuente serían las que no se pueden localizar en un solo sitio, sino que pueden tener varios sitios. Ejemplos de este caso en donde se contamina el agua, son: las fabricas, plantas de tratamiento de aguas, pozos de petróleo, buques petroleros de altamar, etc.

Si pensamos en la calidad del agua, seguramente vendría a nuestra mente la idea del agua "ideal", formada solamente por hidrógeno y oxígeno, es decir, la conocida fórmula: H_2O. Sin embargo, el agua encontrada en estado natural nunca está en estado puro, sino que presenta sustancias disueltas y en suspensión. Estas sustancias pueden limitar, de modo igualmente natural, el tipo de usos del agua. En la naturaleza, el agua adquiere una variedad de elementos orgánicos e inorgánicos, estos se describen en [1]:

Inorgánicos: que pueden ser ambiental:

a) relacionados con la atmósfera (involucra a gases que la conforman),

b) relacionada con los tipos de suelos y sus minerales,

c) relacionada con el agua,

d) relacionada con los contaminántes en (a), (b) y (c) hechos por el hombre.

En su circulación por encima y a través de la corteza terrestre, el agua reacciona con los minerales del suelo y de las rocas, lo que le aporta principalmente sulfatos, cloruros, bicarbonatos de sodio y potasio, y óxidos de calcio y magnesio.

Las actividades humanas aportan una variadad de componentes inorgánicos, que llegan a los cuerpos de agua por escurrimientos o por vertidos directos.

Orgánicos: son aportados por escurrimientos que han estado en contacto con la vegetación decayente, con excremento de animales o con desechos de la vida acuática.

La actividad humana también aporta elementos orgánicos al agua natural ya sea por escurrimiento o por vertidos directos.

La calidad del agua no es un criterio completamente objetivo, pero está socialmente definido y depende del uso que se le piense dar al líquido, por lo que cada uso requiere un determinado estándar de calidad. Así por ejemplo, el estándar de calidad del agua, apta para consumo humano no tendrá los mismos parámetros de calidad que los necesarios en una laguna.

Es natural definir la calidád del agua, como un estado de ésta, caracterizado por su composición físico-química y biológica, en que resulta inocua para la vida, dependiendo de su utilidad biológica.

6.1. Principales contaminantes del agua

Según los libros de ecología y medio ambiente [16, 27], menciona que los principales contaminantes del agua se pueden clasificar de la siguiente manera:

- Aguas residuales y otros residuos que demandan oxígeno (en su mayor parte es materia orgánica, cuya descomposición produce la desoxigenación del agua).

- Agentes infecciosos que inhiben el desarrollo de otras formas de vida, como las bacterias o los pirógenos.

- Nutrientes vegetales que pueden estimular el crecimiento de las plantas acuáticas. Éstas, a su vez, interfieren con los usos a los que se destina el agua y, al descomponerse, agotan el oxígeno disuelto y producen olores desagradables.

- Productos químicos, incluyendo los pesticidas, diversos productos industriales, las sustancias tensioactivas contenidas en los detergentes, y los productos de la descomposición de otros compuestos orgánicos.

- Minerales inorgánicos y compuestos químicos.

- Sedimentos formados por partículas del suelo y minerales arrastrados por las tormentas y escorrentías desde las tierras de cultivo, los suelos sin protección, las explotaciones mineras, las carreteras y los derribos urbanos.

- Sustancias radiactivas procedentes de los residuos producidos por la minería y el refinado del uranio y el torio, las centrales nucleares y el uso industrial, médico y científico de materiales radiactivos.

- El calor también puede ser considerado un contaminante, por ejemplo cuando se usa el agua para la refrigeración de las fábricas y las centrales energéticas, que luego vierten en los ríos.

6.2. Parámetros más comunes

Los parámetros más comúnmente utilizados para establecer la calidad del agua son los siguientes: oxígeno disuelto, pH, sólidos en suspensión, DBO (demanda biológica del oxígeno), DQO (demanda química del oxígeno) y la presencia de fósforo, nitratos, nitritos, amonio, amoniaco, compuestos fenólicos, hidrocarburos derivados del petróleo, cloro residual, zinc total y cobre soluble.

Para una cantidad de contaminantes dada, cuanto mayor sea la cantidad de agua receptora mayor será la dilución de los mismos, y la pérdida de calidad será menor.

Por otra parte, la temperatura tiene relevancia, ya que los procesos de putrefacción y algunas reacciones químicas de degradación de residuos potencialmente tóxicos se pueden ver acelerados por el aumento de la temperatura.

Existen 2 tipos de análisis: cualitativos y cuantitativos.

Análisis Cualitativo: Se determina mediante la descripción de características visibles del agua, incluyendo turbidez, claridad, sabor, color y olor del agua:

- La materia suspendida en el agua absorbe la luz, haciendo que el agua tenga un aspecto nublado. Esto se llama turbidez. La turbidez se puede medir con diversas técnicas, esto demuestra la resistencia a la transmisión de la luz en el agua.

- El sentido del gusto puede detectar concentraciones de algunas décimas a varios centenares de PPM y el gusto puede indicar que los contaminantes están presentes, pero no se puede identificar algunos contaminantes específicos.

- El color puede sugerir que las impurezas orgánicas estén presentes. En algunos casos el color del agua puede ser causado incluso por los iones de metales. El color es medido por la comparación de diversas muestras visualmente o con un espectrómetro.

- La detección del olor puede ser útil, porque el oler puede detectar generalmente incluso niveles bajos de contaminantes. Sin embargo, en la mayoría de los países la detección de contaminantes con olor está limitada a terminantes regulaciones, pues puede ser un peligro para la salud, cuando algunos contaminantes peligrosos están presentes en una muestra.

- La cantidad total de materia suspendida puede ser medida filtrando las muestras a través de una membrana, secando y pesando el residuo.

Análisis cuantitativo: La identificación y la cuantificación de contaminantes disueltos se hace por medio de métodos muy específicos en laboratorios, porque éstos son los contaminantes

que se asocian a riesgos para la salud.

La calidad del agua se puede también determinar por un número de análisis cuantitativos en el laboratorio, tales como pH, sólidos totales (ST), la conductividad y la contaminación microbiana.

- El pH es el valor que determina si una sustancia es ácida, neutra o básica, calculado el número de iones de hidrógeno presentes.

  La escala es a partir de 0 a 14, el medio (7) es neutro. Los valores del pH por debajo de 7 indican que una sustancia es ácida y los valores de pH por encima de 7 indica que es básica.

  Cuando una sustancia es neutra el número de átomos de hidrógeno y de oxhidrilos es igual. Cuando el número de átomos de hidrógeno ($H+$) excede el número de átomos del oxhidrilo ($OH-$), la sustancia es considerada ácida.

  La escala para análisis de pH es la siguiente (Figura 6.1):

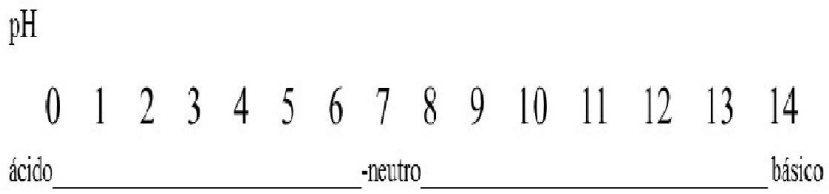

Figura 6.1: Escala de medicion del pH.

- Los sólidos totales (ST) son la suma de todos los sólidos disueltos y suspendidos en el agua.

  Cuando el agua se analiza para los ST se seca la muestra y el residuo se pesa después. ST pueden ser tanto las sustancias orgánicas como inorgánicas, los microorganismos y partículas más grandes como la arena y arcilla.

- La conductividad específica, signífica la conducción de la energía por los iones. La medida de la conductividad del agua puede proporcionar una visión clara de la concentración de iones en el agua, pues el agua es naturalmente resistente a la conducción de la energía.

  La conducción se expresa en Siemens y se mide con un conductivímetro o una célula.

  Las reacciones químicas que tienen lugar durante la realización de esta técnica son las mismas que ocurren durante un proceso de electrólisis, la oxidación en el ánodo y la reducción en el

cátodo.

- La contaminación microbiana es dividida en la contaminación por los organismos que tienen la capacidad de reproducirse, de multiplicarse y los organismos que no pueden hacerlo.

- La contaminación microbiana puede ser la contaminación por las bacterias, que es expresada en Unidades Formadoras de Colonias (UFC), es una medida de la población bacteriana.

- Otra contaminación microbiana es la contaminación por pirógenos. Pirógenos son los productos bacterianos que pueden inducir fiebre en animales de sangre caliente. Después de bacterias y de 4 pirogen las aguas se pueden también contaminar por los virus.

- Los análisis se pueden también hacer por medidas del carbón orgánico total (COT) y por la demanda biológica y química de oxígeno.

- El DBO es una medida de la materia orgánica en el agua, con una unidad de medición mg/l. Es la cantidad de oxígeno disuelto que se requiere para la descomposición de la materia orgánica. La prueba de la DBO toma un período de cinco días.

- El DQO es una medida de la materia orgánica e inorgánica en el agua, con unidades de medición mg/l es la cantidad de oxígeno disuelto requerida para la oxidación química completa de contaminantes.

6.2.1. Base de datos

La base de datos obtenida, fue gracias a la ayuda de la SEMARNAT, la cuál tiene 12 parámetros y 200 observaciones La informacion es anual desde 1990 hasta 2006, su ubicación geográfica se presenta en la Figura 6.2 de los principales ríos de México.

- Río Balsas.
- Río Bravo.
- Río Colorado.
- Río Grijalva.
- Río Lerma.
- Río Jamapa.
- Río Moctezuma.
- Río Pánuco.
- Río Papaloapan.
- Río San Juan.

- Río Soto la Marina.
- Río Tula.

De los cuales, se midieron los siguientes parámetros con su unidad de medición, que se muestra en la Tabla 6.1:

Tabla 6.1: Variables registradas en los ríos con sus unidades de medición.

Parámetro	Unidad de medición
amonio (NH_4)	($mg\ N/l$)
colíformes fecales	($nmp/100ml$)
DBO	(20^oC), ($mg\ O_2/l$)
DQO	($mg\ O_2/l$)
oxígeno disuelto	($mg\ O_2/l$)
sólidos disueltos	(mg/l)
sólidos suspendidos	(mg/l)
pH	(en laboratorio)
conductividad específica	($\mu mhos/cm$)
temperatura	(^oC)
nitratos (NO_3)	($mg\ N/l$)
ortofosfatos	(mg/l)

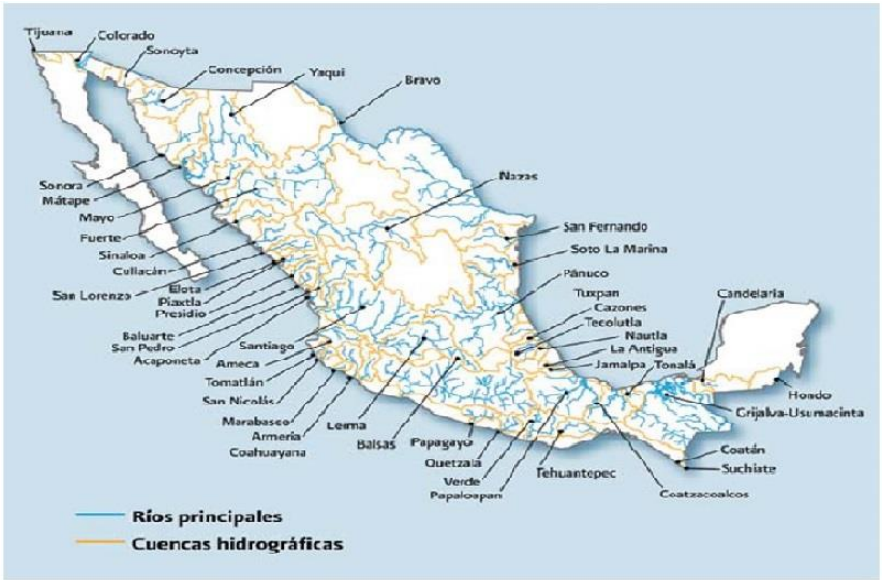

Figura 6.2: Principales ríos de México.

6.2.2. Limpieza de la base de datos

Uno de los principales problemas que se tiene, es la limpieza de la base de datos extraída de la SEMARNAT.

¿Por qué uno tiene que limpiar la base de datos?
Porque muchas veces no están completas, además no nos interesan las observaciones con ceros, dado que no aportarían información significativa a nuestro problema, que en este caso sería la contaminación del agua y así poder emplear un modelo de regresión lineal.

De los cuales, los parámetros sobrevivientes fueron: 10 parámetros de 12, y 169 observaciones de 200; amonio, colíformes fecales, DBO, DQO, oxígeno disuelto, sólidos suspendidos, sólidos disueltos, pH, conductividad específica y temperatura.

6.3. Ajuste bayesiano

El modelo de regresión lineal bayesiano a considerar, para la modelación de la relación entre el parámetro DBO con los siguientes parámetros: amonío, colíformes fecales, DQO, oxígeno disuelto, sólidos súspendidos, sólidos disueltos, pH, conductividad específica, temperatura, es el siguiente.

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

cuya componente sistemática es:

$$\mu = X\beta.$$

El modelo de regresión lineal bayesiano, se caracteriza por precisar una distribución a priori para los parámetros del modelo. Esta distribución, debe poder describir la incertidumbre sobre los verdaderos valores de los parámetros, una vez observados los datos; la información sobre los parámetros estará contenida en la distribución a posteriori.

De lo cual, tenemos que considerar,

a) *Especificar un modelo probabilístico* $p(DBO|\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$. Donde σ^2 es la matriz de covarianzas de tamaño $n \times n$.

$$p(DBO|\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp -\frac{1}{2\sigma^2}[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + n\widehat{\sigma^2}].$$

b) *Especificar una distribución inicial* $p(\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$. Para los parámetros desconocidos ($\beta y \sigma^2$) del modelo probabilístico a considerar, se debe especificar una distribución inicial que describa la incertidumbre que se tiene sobre sus verdaderos valores, recordemos que utilizaremos a $\tau = \frac{1}{\sigma^2}$ como la presición.

- Informativa.
  - Conjugada:

$$p(\beta, \tau) = p(\beta|\tau)p(\tau) = N(\beta|b_0, \tau^{-1}B_0)Ga\left(\tau|\frac{a}{2}, \frac{d}{2}\right).$$

- No informtiva.
  - Inicial de Jeffreys:

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

c) *Determinar la distribución final* $p(\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n|DBO)$.

La determinación de la distribución final o a posteriori se obtiene a través del teorema de Bayes:

$$p(\beta, \tau|DBO) \propto l(DBO|\beta, \tau)p(\beta, \tau).$$

6.3.1. Ajuste bayesiano con a priori informativa

La ecuación de regresión lineal por medio del análisis bayesiano es la siguiente, utilizando una a priori informativa, que en su caso sería la conjugada natural de $p(DBO|\beta, \tau)$.

$$DBO = 20.32 + 2.731\ amonio + 0.00000001999\ colíformes + 0.1178\ DQO$$
$$-2.198\ oxígeno - 0.0007291\ sólidosd + 0.005844\ sólidoss + 1.672\ pH$$
$$+0.003572\ conductividad - 0.8033\ temperatura.$$

$$R^2 = 93\,\%.$$

En la siguiente Tabla 6.2, damos un resumen de los resultados obtenidos del programa, usando una a priori informativa en el modelo de regresión lineal bayesiano, nos muestra las estadísticas de las variables de interés, a través de las cadenas seleccionadas. Nos da los valores de la media, desviación estándar (SD), (MC error), que es una estimación de la desviación estándar por el método de simulación Monte Carlo. Con esto observamos como difieren la media y mediana, de aquí nos damos cuenta de que no se distribuyen como una distribución normal, aunque en algunos parámetros no difiere mucho, como en el caso del oxígeno, también observamos, como on los valores de los coeficientes de los parámetros, especialmente si pueden tomar valores de cero, tales parámetros son: los colíformes fecales, sólidos disueltos, sólidos suspendidos y el pH, estos parámetros pueden tomar valores de cero, porque está contenido dentro de su intervalo de credibilidad.

Tabla 6.2: Resumen estadístico del modelo bayesiano con a priori informativa.

Parámetros	Media	SD	MC errores	Val. 2.5 %	Mediana	Val. 97.5 %
const	20.32	9.839	1.204	1.774	19.2	39.2
amonio	2.731	0.3855	0.01649	1.968	2.733	3.467
colíformes	1.999×10^{-8}	3.2×10^{-8}	1.06×10^{-9}	-4.158×10^{-8}	1.953×10^{-8}	8.39×10^{-8}
DQO	0.1178	0.01365	0.0009085	0.0903	0.1179	0.1446
oxígeno	−2.198	0.4255	0.04134	−3.039	−2.195	−1.381
sólidosd	−0.0007291	0.001844	0.0001089	−0.004259	−0.0007728	0.00297
sólidoss	0.005844	0.009784	0.0002805	−0.01313	0.00562	0.0253
pH	1.672	1.058	0.1274	−0.5409	1.834	3.294
conductividad	0.003572	0.001541	0.0001049	0.0005501	0.0036	0.006541
temperatura	−0.8033	0.2074	0.0232	−1.164	−0.723	−0.3323
τ	0.0191	0.001921	0.00003561	0.01548	0.001906	0.02306
σ^2	2826	586.4	12.0	1882.0	2752.0	4172.0

- Usando la metodología bayesiana, se realizaron 5000 iteraciones, de las cuales, las primeras 1000 se descartaron en cada cadena, para encontrar el proceso de estabilización, en la Figura 6.2.0 se puede apreciar el gráfico del valor del parámetro contra el número de iteraciones, muestra la estabilización y convergencia de las cadenas en cada uno de los parámetros. Esta gráfica es dinámica, es decir, a medida que corren las iteraciones del parámetro. los valores generados aparecen en la gráfica, como puede observarse, los parámetros que mejor se comportan en el programa son: el amonio, colíformes fecales, DQO, sólios suspensidos, incluyendo a σ^2 y τ.

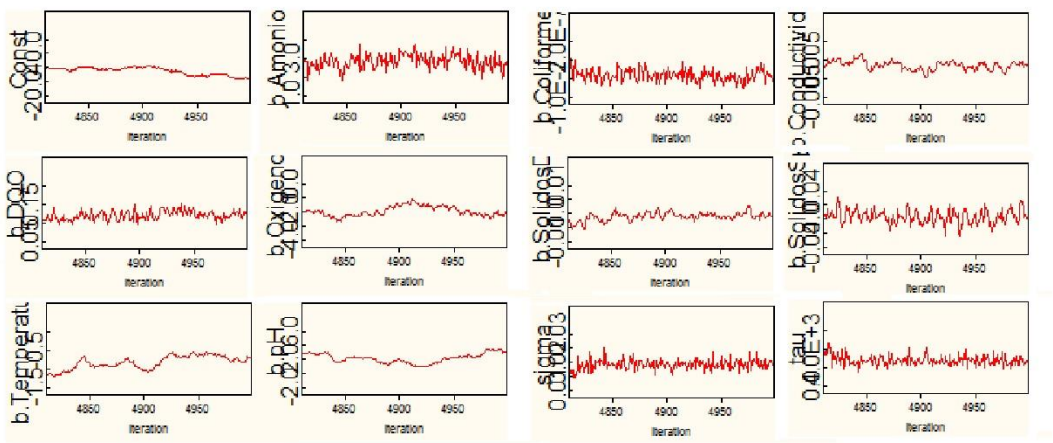

Figura 6.2.0: Iteraciones de los parámetros.

- En la Figura 6.2.1, se muestra las gráficas de los intervalos de credibilidad al 95 % de probabilidad para las medias generadas en cada iteración, como se observa, en cada intervalo, el valor del parámetro cae dentro de cada intervalo de credibilidad, por lo tanto son adecuados para el modelo, .

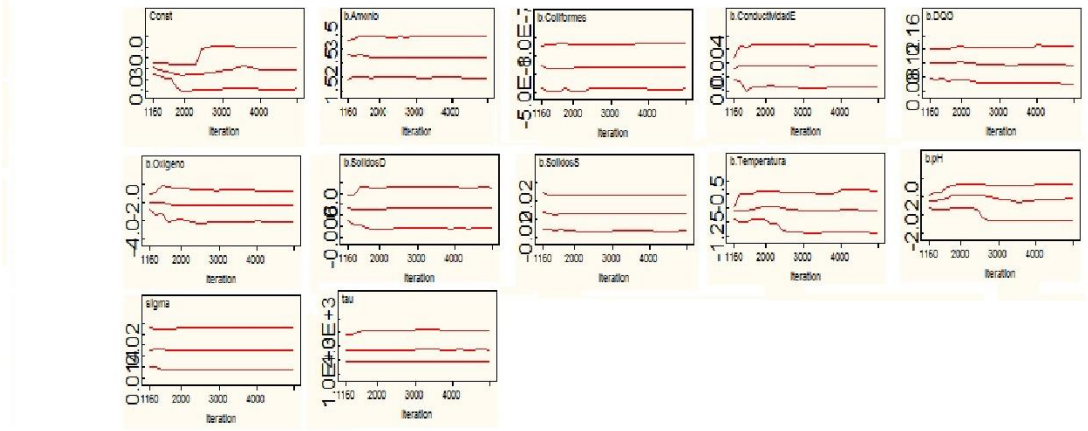

Figura 6.2.1: Gráfico de cuántiles.

- En la Figura 6.2.2, se muestra la autocorrelación entre los valores simulados por cadenas para cada parámetro. Aunque OpenBUGS sólo muestra un rezago (lag) de 50, puede considerarse que los valores por cadena, para cada parámetro son independientes.

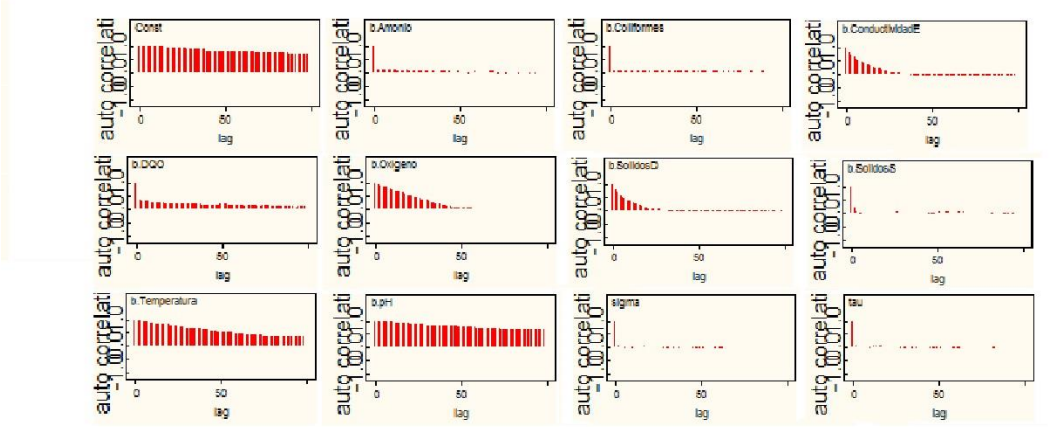

Figura 6.2.2: Autocorrelaciones.

- La distribución pósterior o final, resulta la combinación de las cadenas generadas para cada parámetro, como se observa en la Figura 6.2.3, los parámetros no se comportan como una distribución normal, aunque como nosotros estamos trabajando con varias observaciones y además presentan independencia entre sí (autocorrelación), no nos afecta demasiado en los resultados, pero si nos fijamos en la Tabla 6.2 nos damos cuenta de que la media y la mediana son diferentes, por lo que al momento de graficar su distribución posterior nos da como resultado las siguientes gráficas.

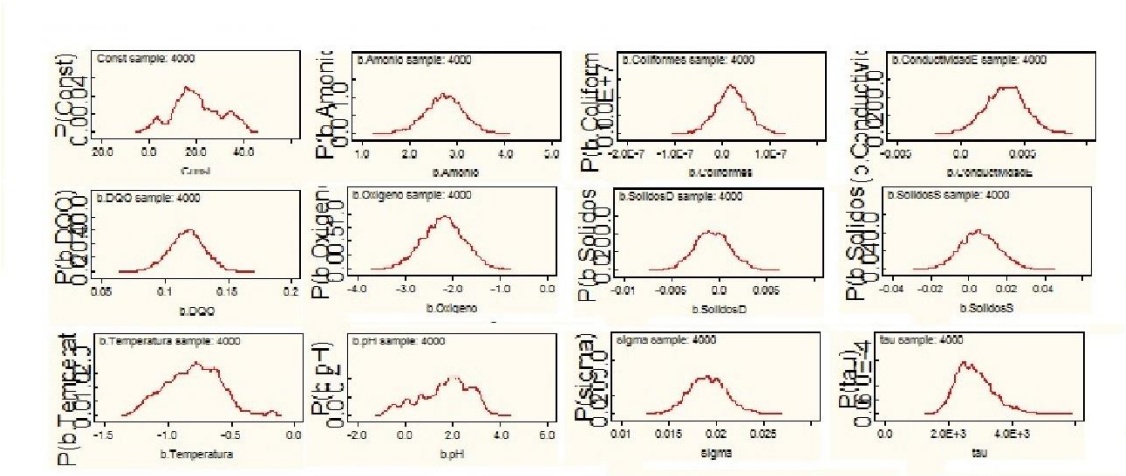

Figura 6.2.3: Función de densidad posterior.

- En la Figura 6.2.4, se muestra una gráfica del parámetro contra las iteraciones, sólo se muestran algunas iteraciones que fueron seleccionadas, de aquí, podemos observar que los parámetros que mejor se comportan en el modelo, son el amonio, colíformes fecales, DQO, conductividad específica, los sólidos suspendidos y disueltos, oxígeno, τ y σ^2.

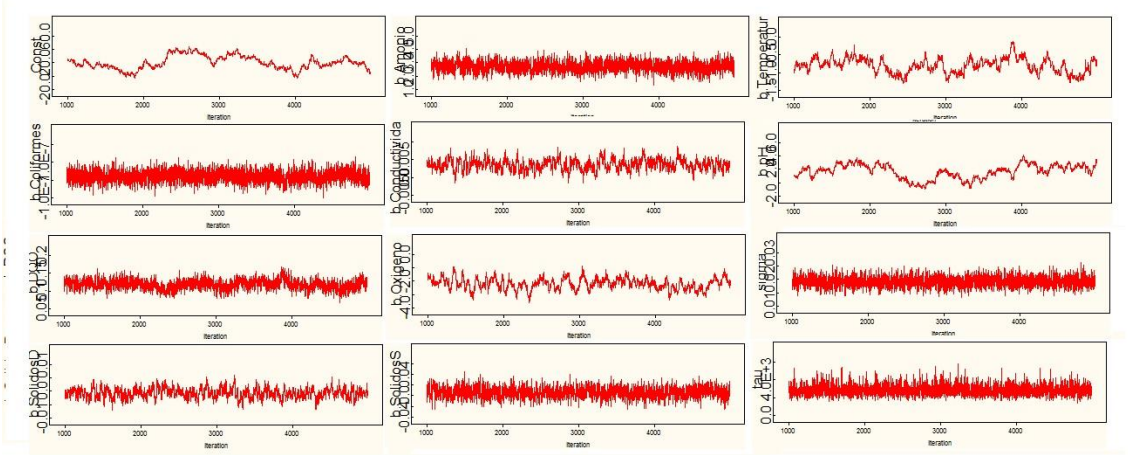

Figura 6.2.4: Histograma.

6.3.2. Ajuste bayesiano con a priori no informativa

En este caso, la ecuación de regresión lineal por medio del análisis bayesiano es la siguiente, utilizando una a priori no informativa, para este caso sería la inicial de Jeffreys, $p(\beta, \tau) \propto \tau$, donde $\tau = \frac{1}{\sigma^2}$.

$$DBO = 21.54 + 2.714\ amonio + 0.00000001727\ colíformes + 0.1187\ DQO$$
$$-2.24\ oxígeno - 0.0006948\ sólidosd + 0.005838\ sólidoss + 1.435\ pH$$
$$+0.003679\ conductividad - 0.7708\ temperatura.$$

$$R^2 = 79\,\%.$$

En la Tabla 6.3, damos un resumen de los resultados obtenidos del programa, cuando usamos una a priori no informativa en el modelo de regresión lineal bayesiano, nos muestra las estadísticas de las variables de interés, a través de las cadenas seleccionadas. Observamos los valores de la media, (SD) desviación estándar, (MC error) una estimación de la desviación estándar Monte Carlo para la media, mediana y los cuantiles al 2.5 % y 9.75 % de probabilidad. Con esto vemos como difieren la media y mediana, por lo que no se comportan como una distribución normal, aunque en algunos parámetros no difiere mucho, como en el caso del oxígeno, aquí también podemos ver

si los parámetros pueden tomar valores de cero, como en el caso anterior, en donde usamos una a priori informativa, los parámetros que pueden tomar valores de cero son: los colíformes fecales, sólidos suspendidos, sólidos disueltos y el pH, que son los mismos parámetros que en el caso anterior.

Tabla 6.3: Resumen estadístico del modelo bayesiano con a priori no informativa.

Parámetros	Media	SD	MC errores	Val. 2.5 %	Mediana	Val. 97.5 %
const	21.54	8.93	1.088	1.23	21.48	37.84
amonio	2.714	0.4024	0.0174	1.918	2.712	3.51
colíformes	1.727×10^{-8}	3.367×10^{-8}	1.064×10^{-9}	-4.9×10^{-8}	1.726×10^{-8}	8.485×10^{-8}
DQO	0.1187	0.01426	0.000985	0.09051	0.1187	0.1472
oxígeno	−2.24	0.4865	0.04713	−3.085	−2.28	−1.117
sólidosd	−0.0006948	0.00202	0.0001153	−0.004607	−0.0006816	0.003377
sólidoss	0.005838	0.01005	0.0002755	−0.01385	0.005877	0.02563
pH	1.435	1.01	0.121	−0.6226	1.533	3.439
conductividad	0.003679	0.001736	0.0001178	0.00005138	0.0037	0.007065
temperatura	−0.7708	0.2331	0.02699	−1.186	−0.7753	−0.2807
τ	0.01751	0.001997	0.00004149	0.01379	0.001744	0.02173
σ^2	57.85	6.679	0.141	46.03	57.36	72.51

- Las iteraciones realizadas por cadena fueron de 5000, de las cuales, las primeras 1000 se descartaron en cada cadena, debido a que se encontraban en un proceso de estabilización, en la Figura 6.3.0 se puede apreciar el gráfico del valor del parámetro contra el número de iteraciones. Esta gráfica es dinámica, aquí se puede observar, los valores que mejor se ajustan al modelo propuesto, que son: el amonio, colíformes fecales, DQO, sólidos suspendidos, se asemeja un poco a los valores obtenidos usando una a priori informativa de la Figura 6.2.0.

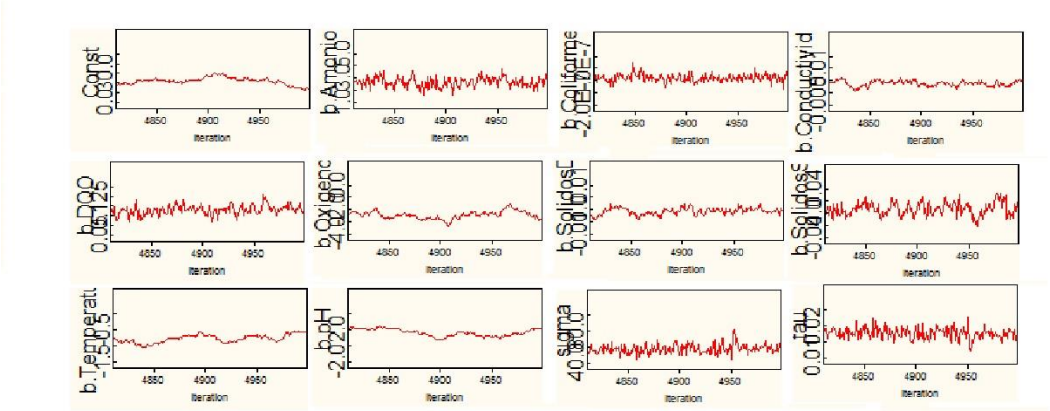

Figura 6.3.0: Iteraciones de los parámetros.

- En la Figura 6.3.1, se muestra las gráficas de los intervalos de credibilidad al 95 % de probabilidad para las medias generadas en cada iteración, se puede observar que los intervalos de credibilidad son buenos, ya que caen dentro del intervalo los valores de mis parámetros.

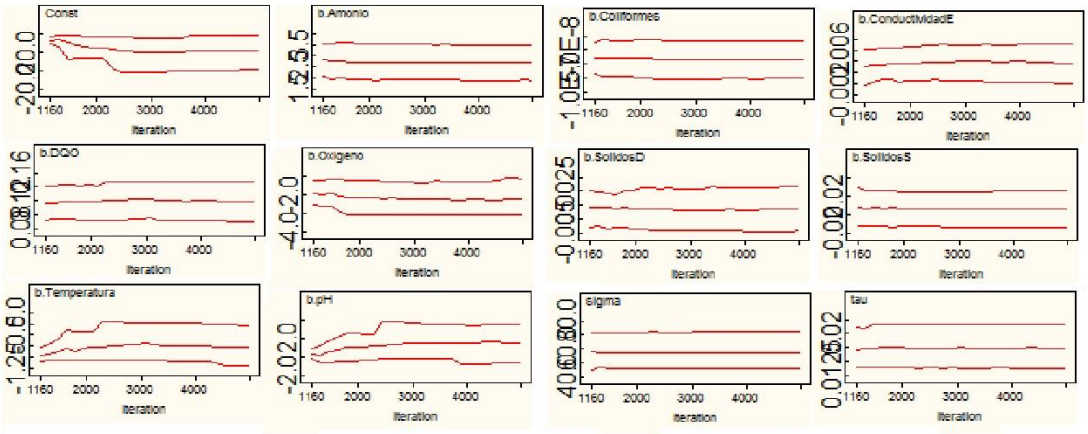

Figura 6.3.1: Gráfico de cuántiles.

- La Figura 6.3.2 de abajo, se muestra la autocorrelación entre los valores simulados por cadena para cada parámetro, puede considerarse que los valores por cadena, para cada parámetro son independientes.

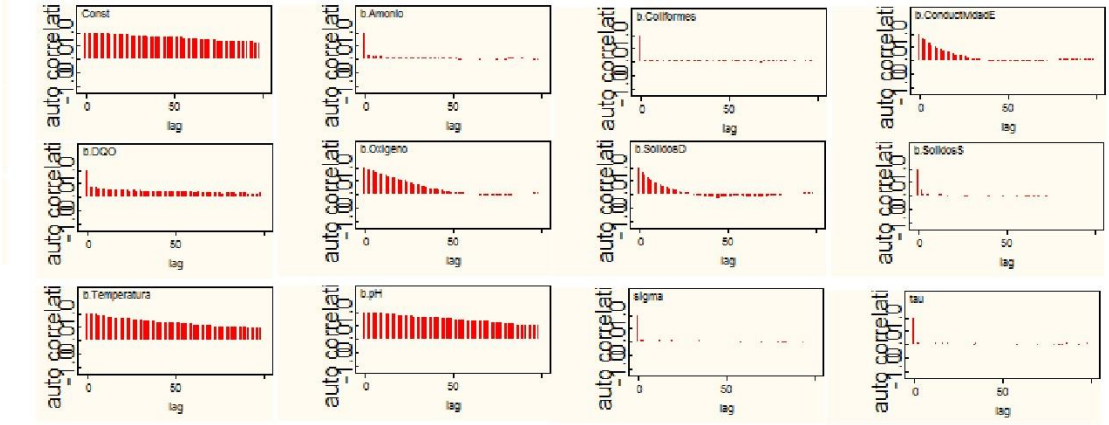

Figura 6.3.2: Autocorrelaciones.

- La Figura 6.3.3 es la gráfica de la distribución posterior o final de cada parámetro, resulta la combinación de las cadenas generadas para cada parámetro, observamos que no se parecen a una distribución normal, aunque el parámetro de conductividad específica su gráfica parece tender a una distribución normal. Nuevamente observamos en la Tabla 6.3 que la media y

la mediana son diferentes, al momento de usar una a priori no informativa la Figura 6.3.3 es diferente a la Figura 6.2.3 cuando usamos una a priori informativa en el modelo.

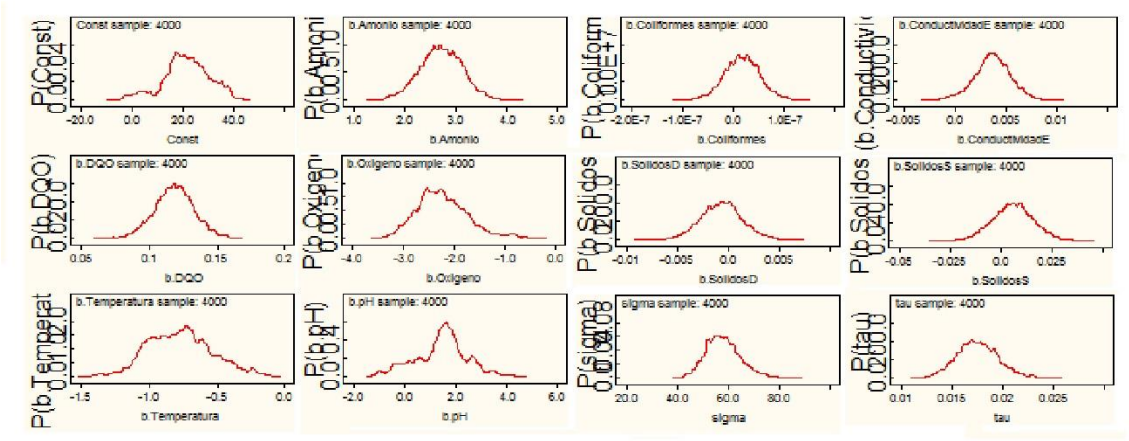

Figura 6.3.3: Función de densidad pósterior.

- En la Figura 6.3.4, se muestra una gráfica del parámetro contra las iteraciones, sólo muestra las iteraciones que fueron seleccionadas, de las cuales, las que mejor se comportan al modelo son: amonio, colíformes fecales, DQO, sólidos suspendidos, conductividad específica, sólidos disueltos, τ y σ^2, el oxígeno no se comporta de manera adecuada, en comparación cuando utilizamos una a priori informativa, pero no hay mucha diferencia entre la Figura 6.3.4 y la Figura 6.2.4.

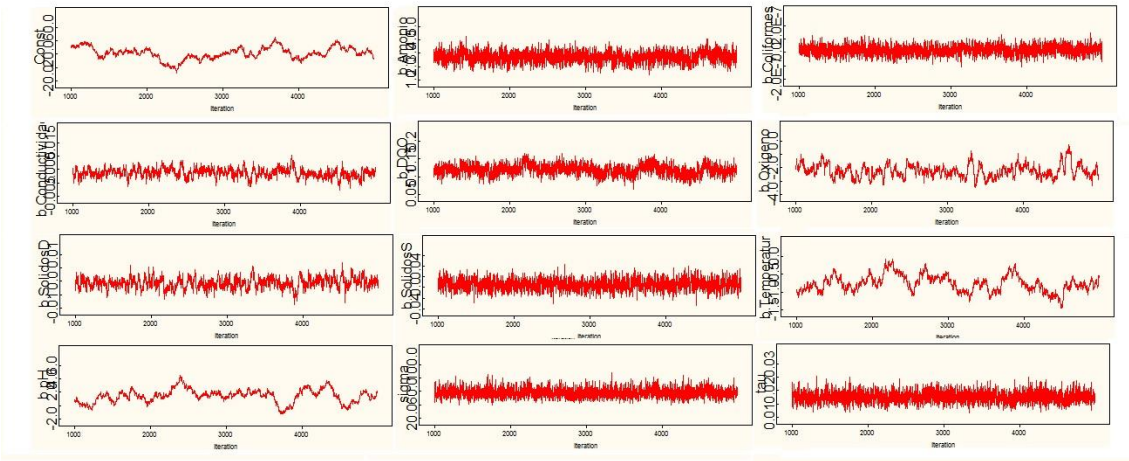

Figura 6.3.4: Histograma.

6.4. Ajuste clásico

La ecuación de regresión multiple usando (MCO) en estadística clásica es:

$$DBO = 16.2 + 2.73\ amonio + 0.00000002\ colíformes + 0.119\ DQO$$
$$-2.23\ oxígeno - 0.00075\ sólidosd + 0.0058\ sólidoss + 2.04\ pH$$
$$+0.00362\ conductividad - 0.749\ temperatura.$$

$$R^2 = 70.3\,\%,\ R^2(ajustado) = 68.6\,\%,\ S = 7.50794.$$

En la Tabla 6.4, se muestra un resumen estadístico del modelo, nos fijamos que el valor crítico P es menor que el valor del estadístico T, por lo que podemos asegurar que los valores de β_i con $i = 0, ..., 9$ no son de cero, con excepción de los colíformes fecales, sólidos disueltos y sólidos súspendidos, en donde el valor crítico P es mayor o casi igual al valor del estadístico T, la diferencia con los otros modelos a parte de los valores que toma cada parámetro es que el valor del pH no puede ser de cero, porque su valor P es menor al de T.

Tabla 6.4: Análisis de los parámetros del modelo.

Parámetros	Coeficientes	Desv. estándar	T	P
const	16.19	13.1	1.24	0.218
amonio	2.7257	0.4027	6.77	0.000
colíformes	0.00000002	0.00000003	0.54	0.587
DQO	0.11916	0.01391	8.57	0.000
oxígeno	−2.2293	0.4577	−4.87	0.000
sólidosd	−0.000755	0.01905	−0.4	0.692
sólidoss	0.00579	0.01	0.58	0.564
pH	2.04	1.415	1.44	0.151
conductividad	0.003618	0.00163	2.22	0.028
temperatura	−7.49	0.2233	−3.35	0.001

El estadístico de $Durbin - Watson = 1.19402$.

Como el estadístico de Durbin-Watson no es muy grande, las tablas conocidas no proporcionan valores críticos para $n = 169$ y $K = 9$, podemos asegurar que no existe autocorrelación entre los parámetros.

Para que se pueda llevar a cabo esta prueba se han elaborado tablas que contienen los puntos $d_{L,\alpha}$ y $d_{U,\alpha}$ apropiados para varios va lores de α (tamaño de la prueba), k (número de variables

independientes) y n (número de observaciones).

En la Tabla 6.5, se muestra el análisis de varianza del modelo en si, por lo que podemos ver es que si hay una relación lineal significativa entre DBO y los demás parámetros.

Tabla 6.5: Análisis de varianza del modelo.

Fuente	Grados de Libertad	Suma de cuadrados	Promedio de los cuadrados	F	P
Regresión	9	21206.3	2356.3	41.80	0.000
Error residual	159	8962.7	56.4		
Total	168	30169.0			

- Prueba de normalidad de los parámetros, utilizando la prueba de Anderson-Darling, utilizando el programa de minitab 16 [22].

  - Los parámetros Amonío y Colíformes fecales, no cumple con la prueba de normalidad porque el estadístico de $A.D.$ es muy grande con respecto al $P-valor$, ver Figura 6.4.0.

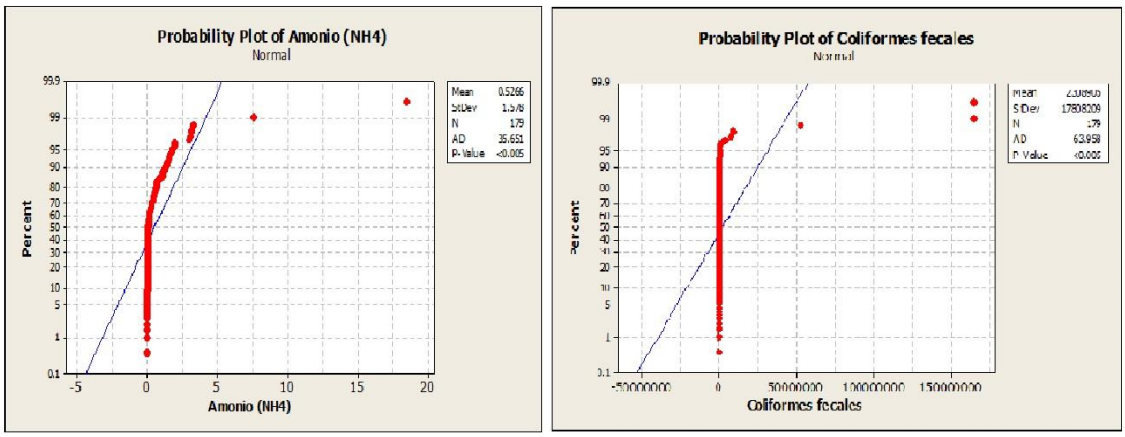

Figura 6.4.0: Prueba de normalidad de los parámetros: Amonío y Colíformes fecales.

  - Para ambos parámetros DBO y DQO el estadístico de $A.D.$ es muy grande con respecto a su $P-Valor$ por eso decimos que no cumple con la prueba de normalidad, Figura 6.4.1.

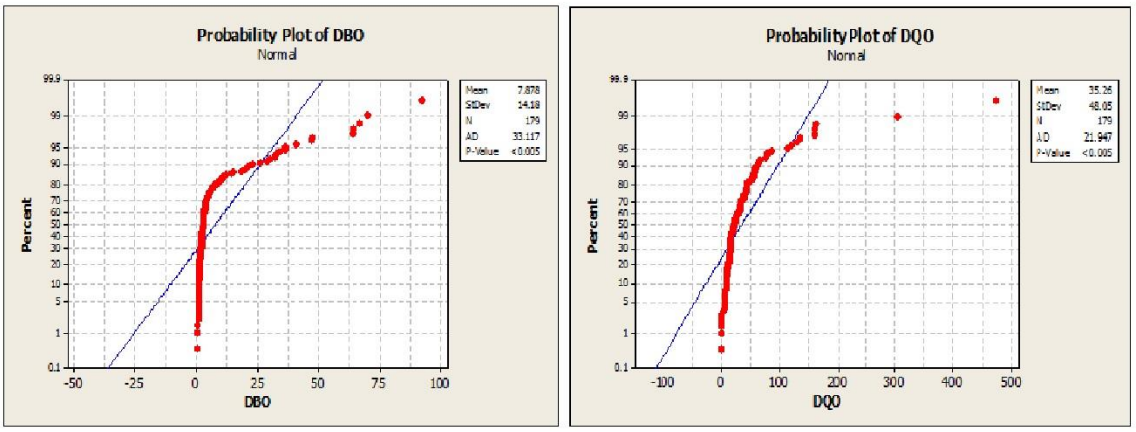

Figura 6.4.1: Prueba de normalidad de los parámetros: DBO y DQO.

- En el caso del Oxigeno disuelto casi parece que tiende a una normal pero no, porque su estadístico de $A.D.$ es algo elevado con respecto a su $P - Valor$, por eso decimos que no puede tener una distribución normal, en el caso de los Sólidos Disueltos, ahí decimos que no se comporta como una normal, Figura 6.4.2.

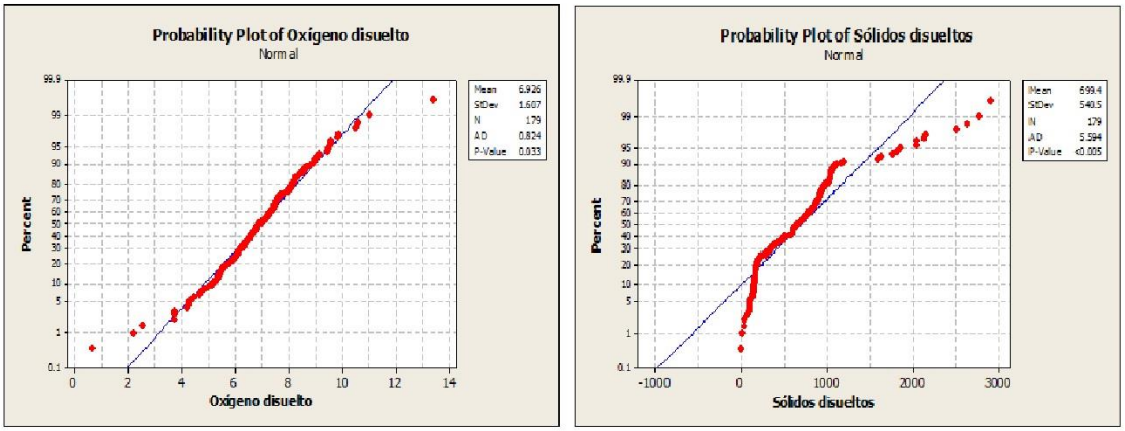

Figura 6.4.2: Prueba de normalidad de los parámetros: Oxigeno disuelto y Sólidos disueltos.

- Es el turno de los Sólidos Suspendidos y el pH, ambos parámetros no tienden a comportarse como una normal, porque el $A.D.$ es muy elevado con respecto de su $P - valor$, Figura 6.4.3.

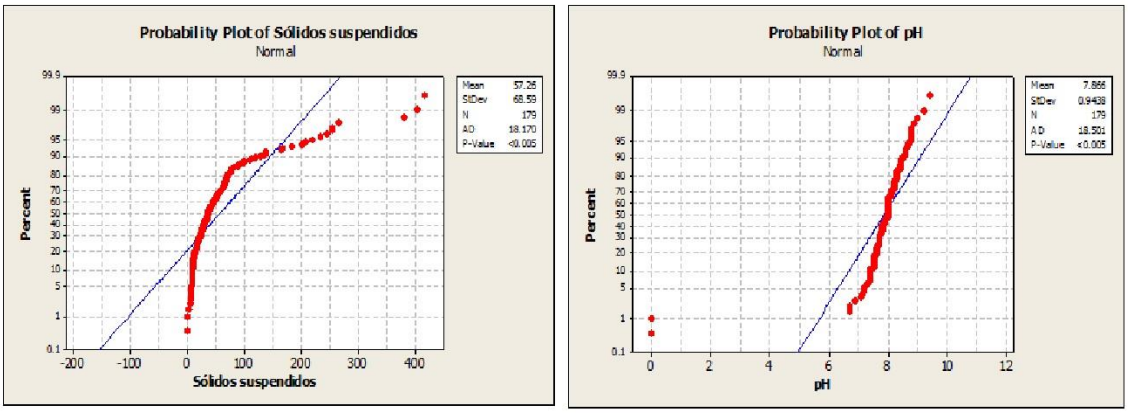

Figura 6.4.3: Prueba de normalidad de los parámetros: Sólidos suspendidos y pH.

- Nos queda probar las dos ultimas variables, que son, la Conductividad Específica y la Temperatura, ambos parámetros no cumplen con la prueba de normalidad, porque ambos tienen el estadístico de $A.D.$ muy elevado en comparación con el $P - valor$, Figura 6.4.4.

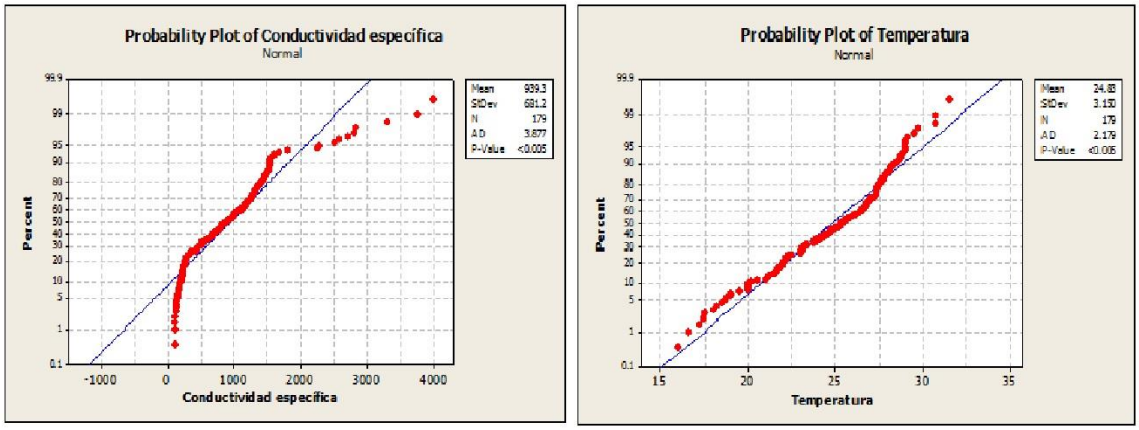

Figura 6.4.4: Prueba de normalidad de los parámetros: Conductividad específica y Temperatura.

- El recuadro, gráficos de residuos tipificados, contiene dos opciones; gráficas que informan sobre el grado en el que los residuos tipificados se aproximan a una normal, veamos si lo cumplen, primero observemos lo siguiente:

- Histograma. Ofrece un histograma de los residuos tipificados con una curva normal superpuesta. La curva se construye tomando una media 0 y una desviación típica de 1, es decir, la misma media y desviación típica que los residuos tipificados.

En el histograma de la figura 6.4.5 de abajo, podemos observar, en primer lugar que la parte central de la distribución acumula muchos más casos de los que existen en una curva normal. En segundo lugar, la distribución es algo asimétrica, especialmente en los valores extremos.

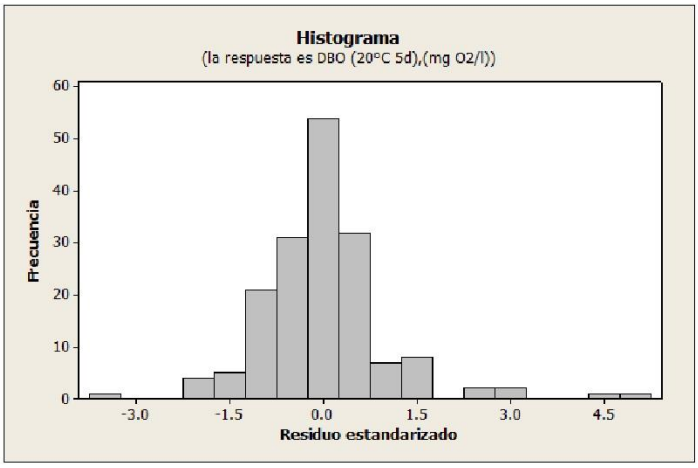

Figura 6.4.5: Gráfico del histograma.

- Gráfico de probabilidad normal. Permite obtener un diagrama normal. En el eje de abcisas está representada la probabilidad acumulada que corresponde a cada residuo tipificado. el de ordenadas representa la probabilidad acumulada teórica que corresponde a cada puntuación típica en una curva normal con media 0 y desviación típica 1. Cuando los residuos se distribuyen normalmente, la nube de puntos se encuentra alineada sobre la diagonal del gráfico (ver figura 6.4.6).

  El gráfico de probabilidad normal, muestra información similar a la ya obtenida con el histograma. Los puntos no se encuentran alineados sobre la diagonal del gráfico, especialmente los puntos extremos.

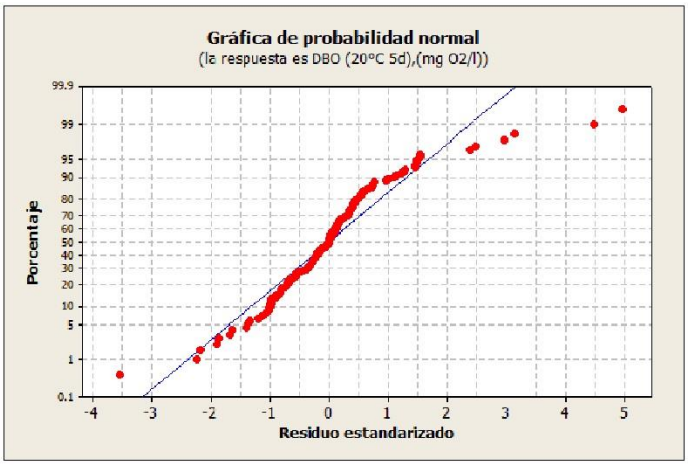

Figura 6.4.6: Gráfico de probabilidad normal.

- La Figura 6.4.7, el plot de los residuales *vs* el índice de la observación muestra que la observación 73 y 100 son conocidos como "*outlier*", pues el residual estandarizado cae más allá de 4.

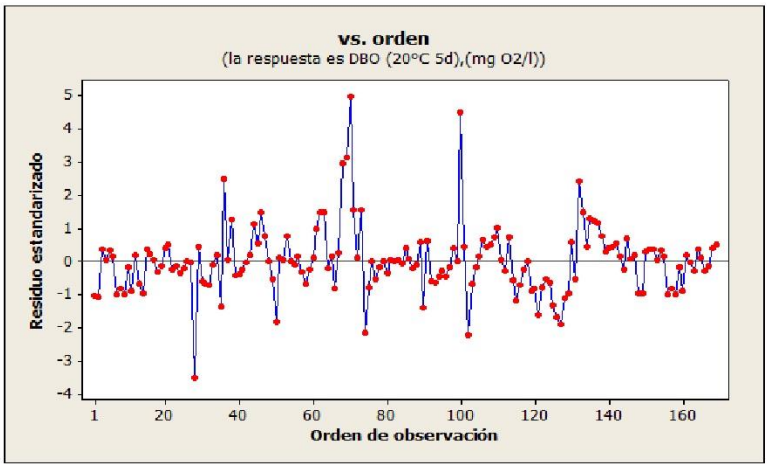

Figura 6.4.7: Gráfica de residuos *vs* orden.

- La Figura 6.4.8, el plot de los residuales *vs* los valores predichos muestra que la varianza de los errores no es constante con respecto a la variable de respuesta, pues tiende ha aumentar cuando el valor de la variable de respuesta aumenta.

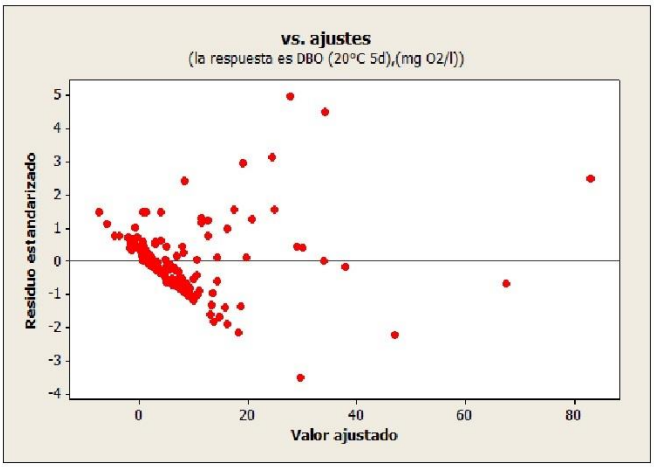

Figura 6.4.8: Gráfica de residuales *vs* valores.

6.4.1. Selección de variables

A continuación se presenta el modelo de regresión lineal usando, las variables que mejor se comportan, este tipo de procedimiento se le conoce como selección de variables, usando solo las mejores variables para el modelo, podemos encontrar la mejor ecuación de regresión.

El resultado que nos dio es la siguiente ecuación,

$$DBO = 32.4 + 2.81\ amonio + 0.120\ DQO$$
$$-2.13\ oxígeno + 0.003343\ conductividad - 0.782\ temperatura.$$

$$R^2 = 69.8\,\%,\ R^2(ajustado) = 68.9\,\%,\ S = 7.47811.$$

En la Tabla 6.6, nos muestra un resumen estadístico con las mejores variables selecionadas del modelo, nos fijamos que el valor crítico P es menor que el valor del estadístico T, por lo que podemos asegurar que nuestros valores de β_i con $i = 0, ..., 5$ no es de cero, como en la tabla 6.5, en donde los valores de los parámetros de colíformes fecales solídos disueltos y sólidos súspendidos pueden ser de cero, con esto, vemos la prueba de significanza de cualquier coeficiente individual de regresión.

Tabla 6.6: Análisis de los parámetros del modelo con varibles selccionadas.

Parámetros	Coeficientes	Desv. estándar	T	P
const	32.413	5.980	5.42	0.000
amonio	2.8091	0.3867	7.26	0.000
DQO	0.12035	0.01361	8.84	0.000
oxígeno	−2.1251	0.4501	−4.72	0.000
conductividad	0.003429	0.001010	3.40	0.001
temperatura	−7.817	0.2113	−3.70	0.000

En la Tabla 6.7, se muestra el análisis de varianza del modelo cuando usamos selección de variables, por lo que podemos ver es que si hay una relación lineal significativa entre DBO y los demás parámetros.

El estadístico de $Durbin - Watson = 1.22123$.

Como el estadístico de Durbin-Watson no es muy grande, las tablas conocidas no proporcionan valores críticos para $n = 169$ y $K = 5$, podemos asegurar que no existe autocorrelación entre los parámetros.

Tabla 6.7: Análisis de varianza del modelo.

Fuente	Grados de Libertad	Suma de cuadrados	Promedio de los cuadrados	F	P
Regresión	5	21053.7	4210.7	75.30	0.000
Error residual	163	9115.3	55.9		
Total	168	30169.0			

Ahora veamos que las variables que fueron selecionadas son: amonio, *DQO*, temperatura, oxígeno disuelto y la conductividad especídica, pero uando usamos regresión bayesiana, por ejemplo con una apriori informativa, en las gráficas dinámicas y de historial (Figura 6.2.0 y 6.2.4), los parámetros que mejor se comportan en el modelo son: amonio, colíformes fecales, DQO, conductividad específica, los sólidos suspendidos y disueltos, oxígeno, ahora usando una a priori no informativa tenemos los siguientes parámetros que se comportan mejor en las gráficas dinámicas y de historial (Figura 6.3.0 y 6.3.4) son: el amonio, colíformes fecales, *DQO*, sólidos suspendidos, conductividad específica, sólidos disueltos.

De esto podemos ver que en regresión bayesiana los parámetros que mejor se comportan cuando usamos una a priori informativa o no informativa son: amonio,colíformes fecales, *DQO*, conductividad específica y sólidos suspendidos, pero cuando usamos selección de variables en regresión lineal clásica tenemos, el amonio, *DQO*, temperatura, oxígeno disuelto, conductividad específica. Al interceptar los parámetros entre estos dos modelos de regresión tenemos los siguientes parámetros, amonio,*DQO* y conductividad específica, que son los que mejor se comportan en el modelo con respecto al *DBO*.

Con esto concluimos que la regresión lineal clásica aporta resultados no muy diferentes con respecto a la regresión bayesiana, claro que depende mucho de las a prioris que se estén utilizando, como en este caso se usaron unas a prioris que no afecte mucho a nuestro modelo.

6.5. Comparación de los modelos de regresion lineal clásico y bayesiano

En la Tabla 6.8, podemos observar las diferencias, entre trabajar con regresión lineal clásica con respecto a la regresión lineal bayesiana, especialmente nos daremos cuenta de los valores en los coeficientes de los parámetros, cuando usamos una a priori no informativa, a priori informativa y cuando usamos mínimos cuadrados, observaremos que las diferencias entre cada modelo, no son demasiadas entre sí, así pues observaremos también la precisión τ, que nos guíaremos de esto, para saber cual es el mejor modelo.

Tabla 6.8: Tabla de comparación de los modelos.

Parámetros	Regresión clásico	Iniciales no informativas	Iniciales informativas
const	16.9	21.54	20.32
amonio	2.7257	2.714	2.731
colíformes	0.00000002	0.00000001727	0.00000001999
DQO	0.11916	0.1187	0.1178
oxígeno	-2.2293	-2.24	-2.198
sólidosd	-0.000755	-0.0006948	-0.0007291
sólidoss	0.00579	0.005838	0.005844
pH	2.040	1.435	1.672
conductividad	0.003618	0.003679	0.003572
temperatura	-0.7490	-0.7708	-0.8033
τ (precisión)	0.0177304	0.01751	0.0191
σ^2	56.4	57.85	2826.0

El valor del coeficiente de la pendiente (β_0) de los modelos ajustados bayesianamente por la distribución conjugada y la inicial de Jeffreys, no son muy diferentes, sin embargo, existe una gran variación entre los estimadores de la presición (τ), el valor más grande del mismo se obtiene cuando se utilizamos la a priori informativa, por lo que el modelo obtenido bajo la distribución informativa conjugada, es el mejor. Además, parece describir mucho mejor la relación entre el DBO y los demás parámetros observados.

Capítulo 7

Conclusiones

El enfoque bayesiano justifica el uso del conocimiento subjetivo del investigador. Así, esta metodología aprovecha todas las fuentes de información: inicial o a priori (investigaciones anteriores, conocimiento subjetivo, empírico y muestral). Cuando no se cuenta con información a priori, la metodología bayesiana y clásica proponen resultados casi similares. En este caso, la diferencia substancial entre ambos métodos está en el análisis y el enfoque del problema. Cuando se usa una distribución a priori, los resultados bayesianos diferirán de los obtenidos por la metodología clásica. Por lo cual se debe ser muy cuidadoso en la seleción de estos. Al contar con más información (a priori, muestral) los estimadores obtenidos con la metodología bayesiana serán más precisos. El peso de la información a priori y muestral en la distribución posterior es directamente proporcional a la cantidad de información con que se cuente en este caso. Así, si se cuenta con información muestral suficiente, la función de verosimilitud dominará a la distribución a priori.

El teorema de Bayes es válido en todas las aplicaciones de la teoría de la probabilidad. Sin embargo, hay una controversia sobre el tipo de probabilidades que emplea. En esencia, los seguidores de la estadística tradicional sólo admiten probabilidades basadas en experimentos repetibles y que tengan una confirmación empírica mientras que los llamados estadísticos bayesianos permiten probabilidades subjetivas. El teorema de Bayes puede servir entonces para indicar cómo debemos modificar nuestras probabilidades subjetivas cuando recibimos información adicional de un experimento. La estadística bayesiana está demostrando su utilidad en ciertas estimaciones basadas en el conocimiento subjetivo a priori y permite revisar esas estimaciones en función de la evidencia.

Se aplico la metodología a un conjunto de datos reales. En el que consistía de 9 variables con una variable dependiente, en donde el parámetro DBO depende de los otros parámetros restantes que son: amonío, colíformes fecales, DQO, oxígeno disuelto, sólidos suspendidos, sólidos disueltos, pH, conductividad específica y temperatura. Para saber cuál parámetro depende más, se hizó una tabla de correlaciones entre los parámetros y se observó, que la mayoría depende del DBO. Debido a que este trabajo se hizó solamente pensando en estadística bayesiana, no escribimos la matriz de autocorrelación entre los parámetros, pero los resultados se hicieron usando minitab [22]. Se decidió usar un modelo de regresión lineal clásico, obteniendo con ello, un coeficiente de determinación de $R^2 = 70.3\,\%$, con variables que se medían anualmente.

Con un ajuste bayesiano en nuestro modelo, los coeficientes de los parámetros no diferían mucho del todo, especialmente con una inicial no informativa, pero si el $R^2 = 79\,\%$, dado que en la teoría no deben cambiar mucho, pero como se hicieron 5000 iteraciones, de las cuales las primeras 1000 se descartan automáticamente para poder estabilizar el modelo, mis coeficientes cambian un poco, especialmente el del parámetro interceptor, de lo cual, nos da una presición $\tau = 0.01751$,

mientras que en regresión clásica nos da la precisión de, $\tau = 0.0177304$, observamos de que no difiere mucho.

Cuando se hace el ajuste bayesiano, pero con una inicial informativa, los valores de los coeficientes no cambian mucho, en comparación en el modelo no informativo. En especial el valor del regresor interceptor se parece mucho al del regresor con inicial no informativa, $\beta_0 = 21.54 \approx \beta_0 = 20.32$, pero muy diferentes al de regresión clásica $\beta_0 = 16.9$, de lo cuál me dio un $R^2 = 93\,\%$, esto es algo bueno, ya que el coeficiente de determinación aumento de un $R^2 = 70.3\,\%$ a un $R^2 = 93\,\%$, aunque el coeficiente de determinación, no me dice que tan bueno puede ser el modelo, es por eso que no debemos guiarnos mucho por ese lado, es mejor guiarse por la precisión de mis modelos.

El valor más grande de la precisión, se obtiene cuando se ha utilizado la distribución inicial informativa, por lo que se puede considerar que el modelo obtenido bajo la distribución informativa conjugada, es el mejor o el más apropiado para este caso, no obstante, el valor la precisión usando regresión lineal clásica es más grande que cuando usamos el ajuste bayesiano con una a priori no informativa, por lo que también podemos decir que es bueno, pero no el mejor.

De todo esto concluimos, que al hacer uso de la regresión bayesiana, da mejores resultados que al hacer regresión lineal clásica, pero recordemos, que la regresión bayesiana no es muy facíl de aplicar, especialmente si no se tiene un concimiento previo acerca de la estadística bayesiana.

Apéndice A

Código del Programa en OpenBUGS

A.1. Código del programa de regresión lineal bayesiano con iniciales informativas

```
model;
{
   for( i in 1 : N ) {
      DBO[i] ~ dnorm(media[i],sigma);
   }
   for( i in 1 : N ) {
  media[i]<-Const+b.Amonio*Amonio[i]+b.Coliformes*Coliformes[i]+b.Oxigeno*Oxigeno[i]
  +b.DQO*DQO[i]+b.SolidosD*SolidosD[i]+b.SolidosS*SolidosS[i]+b.pH*pH[i]
  +b.ConductividadE*ConductividadE[i]+b.Temperatura*Temperatura[i];
   }

   Const ~ dnorm(0.0,1.0E-6); #### valores que se asignan a los hiperparámetros
   b.Amonio ~ dnorm(0.0,1.0E-6);
   b.Coliformes ~ dnorm(0.0,1.0E-8);
   b.Oxigeno ~ dnorm(0.0,1.0E-6);
   b.DQO ~ dnorm(0.0,1.0E-6);
   b.SolidosD ~ dnorm(0.0,1.0E-6);    ####a prioris informativas
   b.SolidosS ~ dnorm(0.0,1.0E-6);
   b.pH ~ dnorm(0.0,1.0E-6);
   b.ConductividadE ~ dnorm(0.0,1.0E-6);
   b.Temperatura ~ dnorm(0.0,1.0E-6);
   tau~ dgamma(0.001,0.001);
   sigma <-1.0/sqrt(tau);

}


##DATOS

list(N=169)
```

```
rios

##Valores iniciales del algoritmo MCMC
list(Const=0, b.Amonio=0, b.Coliformes=0, b.DQO=0, b.Oxigeno=0, b.SolidosD=0,
b.SolidosS=0, b.pH=0, b.ConductividadE=0, b.Temperatura=0, tau=0.001)
```

La seleccion de $\tau = 0.001$ es por un ejemplo del Doctor Vaquera [31], gracias a esto se obtuvieron buenos resultados.

Dado éstos valores para los hiperparámetros, la distribución conjugada está representando un estado de información inicial mínimo informativo sobre los parámetros (β, τ).

A.2. Código del programa de regresión lineal bayesiano con iniciales no informativas

```
model;
{
   for( i in 1 : N ) {
      DBO[i] ~ dnorm(media[i],tau);
   }
   for( i in 1 : N ) {
  media[i]<-Const+b.Amonio*Amonio[i]+b.Coliformes*Coliformes[i]+b.Oxigeno*Oxigeno[i]
  +b.DQO*DQO[i]+b.SolidosD*SolidosD[i]+b.SolidosS*SolidosS[i]+b.pH*pH[i]
  +b.ConductividadE*ConductividadE[i]+b.Temperatura*Temperatura[i];
   }

   Const ~ dnorm(0.0,1.0E-6);
   b.Amonio ~ dnorm(0.0,1.0E-6);              ####valores para los hiperparámetros
   b.Coliformes ~ dnorm(0.0,1.0E-8);
   b.Oxigeno ~ dnorm(0.0,1.0E-6);
   b.DQO ~ dnorm(0.0,1.0E-6);                       ####a prioris no informativas
   b.SolidosD ~ dnorm(0.0,1.0E-6);
   b.SolidosS ~ dnorm(0.0,1.0E-6);
   b.pH ~ dnorm(0.0,1.0E-6);
   b.ConductividadE ~ dnorm(0.0,1.0E-6);
   b.Temperatura ~ dnorm(0.0,1.0E-6);
   sigma~ dunif(0,1000);
   tau <-1.0/sqrt(sigma*sigma);

}

##Datos

list(N=169)

rios

##Valores iniciales del algoritmo MCMC
```

```
list(Const=0, b.Amonio=0, b.Coliformes=0, b.DQO=0, b.Oxigeno=0, b.SolidosD=0,
b.SolidosS=0, b.pH=0, b.ConductividadE=0, b.Temperatura=0, sigma=0.001)
```

Apéndice B

Métodos de Simulación

La mayor parte del problema de procedimientos bayesianos, se encuentra en poder resolver el cálculo de ciertas características (media, mediana, moda) de la distribución posterior del parámetro de interés [28]. Desafortunadamente, en la práctica la $p(\theta|y)$ puede ser tan compleja que, generalmente, las integrales que implican, pueden no resolverse a mano, así haciendo uso necesario de métodos numéricos eficientes que permitan calcular o aproximar dichas integrales.

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-Hasting 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 Markov (MCMC) permiten generar, de manera iterativa, observaciones de distribuciones multivariadas que difícilmente podrían simularse utilizando 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 se obtiene el 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 [7].*

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, 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.

B.1. Método de Monte Carlo

El método de 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 poder ilustrar el método, suponga que se desea calcular la integral de una función $f(y)$, continua en el intervalo $[a, b]$; es decir:

$$\mathbf{I} = \int_a^b f(y)\, dy.$$

Para poder encontrar un valor aproximado a $\mathbf{I}$ se requiere de una función de densidad $g(y)$, definida en $[a, b]$. Entonces

$$\int_a^b f(y)dy \;=\; \int_a^b f(y)\frac{g(y)}{g(y)}dy.$$

Através de esta expresión se reconoce el valor esperado de la función $h(y) = \frac{f(y)}{g(y)}$, por medio del cuál se obtiene el valor de $\mathbf{I}$, es decir:

$$\mathbf{I} = \int_a^b f(y)\frac{g(y)}{g(y)}dy \;=\; \int_a^b h(y)g(y)d(y) \;=\; E[h(y)].$$

Para obtener este valor esperado es posible simular variables aleatorias independientes e idénticamente distribuidas $Y_1, Y_2, ..., Y_n$, con función de densidad $g(y)$. Por la ley fuerte de los grandes números [21], se tiene

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

Por lo tanto,

$$\hat{\mathbf{I}} = \frac{1}{n}\sum_{i=1}^{n} h(y_i),$$

es un estimador consistente e insesgado de $\mathbf{I}$.

B.2. Métodos computacionales vía cadenas de Markov

Una cadena de Markov es un proceso estocástico $\theta^t : \ t \in T$ tal que, dado el presente, el pasado y el futuro del proceso son independientes [28]. En otras palabras

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

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:

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

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

c) Aperiódica.

d) 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, (ver [28]).

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

a) $\theta^{(t)} \to \theta$, donde $\theta \sim p(\theta|y)$.

b) $\frac{1}{t}\sum_{i=1}^{t} g(\theta^{(i)}) \to E[g(\theta|y)]$, 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 iteracciones 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.

B.3. 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:

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

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 prática, 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:

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

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

c) 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 $p(\theta|y)$ a través de un cociente, por lo que la constante de normalización no es necesaria. 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

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

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

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

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

Es común emplear, después de una reparametrización apropiada, distribuciones de transición normales o t de Student ligeramente sobredispersas, porque se adaptan mejor al problema, es decir el grado en que las observaciones se distribuyen (o separan), ya que nos proporcionan información adicional que nos permite juzgar la confiabilidad de nuestra medida de tendencia central. Si los datos están muy dispersos de la posición central, la medida de tendencia central es menos representativa para los datos, como cuando estos se agrupan más estrechamente alrededor de la media.

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

$$Q_0(\theta^*) = N_d(\theta^*|\hat{\theta}, kV(\hat{\theta}))\ \ (Independencia),$$

donde $\hat{\theta}$ y $V(\hat{\theta})$ denotan a la media y a la matriz de varianzas-covarianzas de la aproximación asintótica normal para $p(\theta|y)$, respectivamente, y $k \geq 1$ es un factor de sobredispersión, porque se ajusta mejor a las características de los datos.

B.4. Algoritmo de Gibbs

El algoritmo de Gibbs permite simular una cadena de Markov $\theta^{(1)}, \theta^{(2)}, ...$ con distribución de equilibrio $p(\theta|y)$. Cada valor nuevo de la cadena se obtiene al generar muestras de distribuciones

cuya dimención es menor que d y que en la mayoría de los casos tiene una forma más sencilla que la de $p(\theta|y)$.

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*:

$$p(\theta_1|\theta_2, ..., \theta_n, y)$$

.

.

.

$$p(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_n, y) \ \ para \ \ toda \ \ i = 2, ..., n-1$$

.

.

.

$$p(\theta_n|\theta_1, ..., \theta_{n-1}, y).$$

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

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

donde $p(\theta|y)$ 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 $p(\theta_1|\theta_2^{(t)}, ..., \theta_n^{(t)}, y)$;

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

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

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

.

.

.

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

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

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

Apéndice C

Distribuciones Iniciales y Posteriores de Referencia

En esta sección se pueden ver en la tabla de abajo, algunos ejemplos de probabilidades posteriores usando probabilidades iniciales conjugadas. Son útiles, porque a veces cuesta trabajo ver que tipo de distribuciones tienden los parámetros cuando estos son tomados como variables aleatorias.

Muéstreo de una población ($\underline{X}$ muéstra aleatoria de tamaño n).	Inicial de referencia	Pósterior de Referencia
Bernoulli (w).	$\prod_{i=1}^{n}(w) \propto w^{-\frac{1}{2}}(1-w)^{-\frac{1}{2}}$	$Beta(\sum_{i=1}^{n} x_i + \frac{1}{2}, n - \sum_{i=1}^{n} x_i + \frac{1}{2})$
Binomial (r, w), r entero conocido.	$\prod_{i=1}^{n}(w) \propto (w)^{-\frac{1}{2}}(1-w)^{-\frac{1}{2}}$	$Beta(\sum_{i=1}^{n} x_i + \frac{1}{2}, rn - \sum_{i=1}^{n} x_i + \frac{1}{2})$
Binomial negativa (r, w), r entero conocido.	$\prod_{i=1}^{n}(w) \propto (w)^{-1}(1-w)^{-\frac{1}{2}}$	$Beta(\sum_{i=1}^{n} x_i, rn - \sum_{i=1}^{n} x_i + \frac{1}{2})$
Poisson (w).	$\prod_{i=1}^{n}(w) \propto w^{-\frac{1}{2}}$	$Gamma(\sum_{i=1}^{n} x_i + \frac{1}{2}, n)$
Exponencial (w).	$\prod_{i=1}^{n}(w) \propto w^{-1}$	$Gamma(n, \sum_{i=1}^{n} x_i)$
Normal (w, r), r (precisión conocida).	$\prod_{i=1}^{n}(w) \propto constante$	$Normal(\overline{x}, nr)$ donde $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.
Normal (m, w), m (media conocida).	$\prod_{i=1}^{n}(w) \propto w^{-1}$	$Gamma(\frac{n}{2}, \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{2})$
Normal (M, w), M media desconocida y w precisión desconocida. Sólo interesa uno de los parámetros.	$\prod_{i=1}^{n}(M, w) \propto w^{-1}$	$N.G.(1, \overline{x}, n, \frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2)$
Normal (M, w), M media desconocida y w precisión desconocida. Interesan los dos parámetros.	$\prod_{i=1}^{n}(M, w) \propto w^{-\frac{1}{2}}$	$N.G.(1, \overline{x}, n, \frac{n+2}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2)$
Multinominal (k, w) donde: K conocido, $w = (w_1, ..., w_k)$ desconocido.	$\prod_{i=1}^{n}(\underline{w}) \propto \prod_{i=1}^{k} w_i^{-\frac{1}{2}}$	$Dirichlet(\underline{\alpha})$ donde $\underline{\alpha} = (y_1 + \frac{1}{2}, ..., y_k + \frac{1}{2})$

Bibliografía

[1] Barba Ho Luz Edith. (2002). Conceptos básicos de la contaminación del agua y parámetros de medición. Universidad del Valle. Facultad de ingeniería de recursos naturales y del medio ambiente, Santiago de Cali.

[2] Berger J. (1980). Statistical decision theory and bayesian analysis. Springer-Verlag: New York, second edition.

[3] Bernardo José M. (2005). Intrinsic credible regions. An objetive bayesian approach to interval estimation. Test 2005; 14(2): 317 − 384 (disponible en http://www.uv.es/ bernardo/2005Test.pdf).

[4] Bernardo José M. and Smith. A.F.M. (1994). Bayesian Theory. Wiley.

[5] Bernardo José M. (2005). Reference analysis. Handbook of statistics 25 (D. K. Dey and C. R. Rao eds.). Amsterdam: Elsevier, 17 − 90.

[6] Bernardo, J. M. (1981). Bioestadística: Una perspectiva Bayesiana. Vicens-Vives.

[7] Cruz, M. E. (2006). WinBUGS: un software para inferencia bayesiana. Tesis de licenciatura, UNAM, México.

[8] Fernández Barberies Gabriela Mónica, María del Carmen Escribabo Ródenas. (2002). El análisis de la robustez y la ayuda a la decisión multicriterio discreta. Departamento de métodos cuantitativos para la economía, facultad de ciencias económicas y empresariales, Universidad San Pablo - CEU, pp. 126 − 138.

[9] French, S. (1986) Decision Theory: An introduction to the mathematics of rationality, Ellis Horwood, Chichester.

[10] Gelman, A. (2009). Bayes, Jeffreys, priori distributions, and the philosophy of statistics. Statistical science 24, 176 − 178.

[11] Gómez Villegas, M.A. (2005). Inferencia estadística. Madrid: Díaz de Santos.

[12] Gómez Villegas Miguel A. Origen de la teoría de la probabilidad. Universidad Complutense de Madrid.

[13] Jeffreys, H. (1961). Theory of probability. Tercera edición. Londres: Oxford University Press.

[14] Kadane, J. B. (1984). Robustness of bayesian analyses, Amsterdam: North-Holland.

[15] Koop, G. (2003). Bayesian Econometrics. New York: John Wiley.& Sons.

[16] Hoeflich Ernesto, Cano Gerónimo, Garza Cuevas Antonio y Martínez Enrique. (1997). Ciencia ambiental y desarrollo sostenible, editorial internacional Thomson, México.

[17] Landaverde Vaca Jóse Guadalupe. (2005). Estimadores bayesianos en la distribución de valores extremos tipo Gumbel, tesis de maestría, Colegio de Postgraduados, Estado de México.

[18] Lee P. M. (1997). Bayesian Statistics An Introduction, Arnold Londres.

[19] Lindley D. V. (1993). The analysis of experimental data. The appreciation of tea and wine. Teaching Statistics; $15 : 22 - 5$.

[20] Matthews R. A. (2000). Facts versus Factions: the use and abuse of subjectivity in scientific research. European Science and Environment Forum Working Paper; reprinted in Rethinking Risk and the Precautionary Principle. Ed: Morris, J. Oxford: Butterworth.

[21] Meyer Paul L. (1999). Probabilidad y aplicaciones estadísticas. Departamento de matemáticas, Washington State University. Addison-Wesley Iberoamericana.

[22] Minitab 16 (2010). Software de análisis estadístico clásico.

[23] Open BUGS. Software para análisis estadístico bayesiano.

[24] Poirier, D. (1995). Intermediate Statistics and Econometrics: A Comparative Approach. Cambridge: The MIT Press.

[25] Press, S. J. (1989). Bayesian Statistics: Principles, Models and Applications. New York: Wiley.

[26] Reyes Cervantes Hortensia. (1987). Selección de variables en análisis de regresión bayesiano. Tesis de licenciatura UNAM, México.

[27] Ricklefs Robert E. (1998). Invitación a la ecología, editorial paramericana, cuarta edición, Buenos Aires.

[28] Robert, C. P. (2007). The bayesian choice: from descison-theoretic foundations to computational implementation. Springer.

[29] www.semarnat.com.mx

[30] Silva LC, Benavides A. El enfoque bayesiano: otra manera de inferir. Gac Sanit 2001; 15:341-6.

[31] Vaquera Huerta Humberto. (2011). Análisis bayesiano con WinBUGS y SAS. Colegio de Postgraduados en ciencias agricolas.

[32] Yupanqui Pacheco R.M. (2005). Introducción a la estadística bayesiana. UNMSM. Facultad de Ciencias Matemáticas. EAP de Estadística, Lima, Perú.