UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ANÁLISIS BAYESIANO DEL MODELO DE REGRESIÓN LINEAL

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I O

P R E S E N T A :

MARCO ANTONIO PÉREZ CASTELLANOS

DIRECTOR DE TESIS:
DRA. RUTH SELENE FUENTES GARCÍA

2016

CD.MX.

1. Datos del alumno
Pérez
Castellanos
Marco Antonio
50 34 33 03
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
309263724

2. Datos del tutor
Dra.
Ruth Selene
Fuentes
García

3. Datos del sinodal 1
Dr.
Carlos
Díaz
Ávalos

4. Datos del sinodal 2
Dra.
Lizbeth
Naranjo
Albarrán

4. Datos del sinodal 2
Dr.
Ricardo
Ramírez
Aldana

5. Datos del sinodal 3
Act.
Jaime
Vázquez
Alamilla

7. Datos del trabajo escrito.
Análisis Bayesiano del modelo de regresión lineal
97p
2016

Ad astra per aspera

Agradecimientos

Doy gracias a mi mamá Esperanza ya que con su enfermedad me ayudó a darme cuenta de todo lo que soy capaz, la vida me ha dado la oportunidad de cuidarla, siempre me ha guiado por el buen camino y ha estado conmigo en los momentos más difíciles, sin duda ella es lo mejor que me ha pasado.

Agradezco a mis hermanos, Roberto por su apoyo incondicional, por todos los sacrificios que hizo por darme lo mejor y por creer siempre en mí, a Francisca por estar siempre conmigo y alegrarme las tardes y la vida. A mi hermana Mayra y a mi tía Francisca, por su amor, por siempre darme ánimos, por su paciencia y por cuidar a mamá mientras yo tomaba clases.

También doy las gracias al destino por poner en mi camino a Arenita, quien me salvó en todo sentido en que se puede salvar a alguien y por darme a Foforita, ambas siempre han estado para mí y son dos grandes motores en mi vida.

Gracias a las personas que estuvieron a mi lado en este camino, a Irlanda, que ha estado conmigo tantos años y a quien ahora extraño tanto, a Gaby por siempre escucharme y darme consejos. A Gloria, Vero y Nayeli que me apoyaron cuando pasaba por un mal momento y nunca me dejaron solo. A Mariana y Antonio que hicieron de mi último año de carrera lo mejor, por ser personas increíbles y darme su amistad sincera. A Cynthia por todos los buenos momentos y su apoyo.

A Joss, por ser mi primera amiga en la vida laboral, por ser una persona increíble en todo aspecto y que siempre tiene una sonrisa dibujada en el rostro. A Ángel, Gabriel y Ricardo por hacer divertido cada día. A May, a quien tanto admiro y que siempre me ayuda.

A la Dra. Ruth por asesorarme en esta tesis, por su paciencia y por todos los consejos que me ha dado y que me han hecho ver que después de todo, las cosas son más simples de lo que parecen.

A todos aquellos que me han brindado una sonrisa o un abrazo, a quienes me han hecho feliz les doy las gracias.

Agradezco a Dios, al Dios infinito que hizo posible esta sucesión de eventos que me llevaron a este momento.

Con esta tesis cierro el ciclo más importante de mi vida, lleno de sueños e ilusiones, me enorgullece mucho ser egresado de mi amada Facultad de Ciencias el lugar donde me formé como profesionista y como persona.

Índice

1. Introducción

La Estadística, puede definirse como un conjunto de técnicas cuyo propósito es la descripción de fenómenos que se manifiestan a través de datos que presentan variabilidad. El método de la estadística se basa en el proceso científico, el cual es básicamente inductivo, que va hasta cierto punto en el orden opuesto al deductivo; este método parte de un proceso de observación, generación de hipótesis, experimentación y potencialmente el pronóstico sobre el comportamiento del fenómeno bajo interés (una finalidad adicional podría ser el establecimiento de leyes o incluso de teorías, de validez general en amplios campos de aplicación).

De este modo, una distinción clave entre la probabilidad y la estadística es que la primera usa el método deductivo, mientras que la segunda es un campo de estudio fáctico y experimental, y se basa en un proceso inductivo, el cual debe de contrastarse en todo caso con la experiencia o la experimentación. Esta definición delimita el ámbito de acción de la disciplina -los fenómenos que presentan variabilidad- y al mismo tiempo, establece su objetivo último: la descripción completa. Así, toda la Estadística es descriptiva y, en particular, la Inferencia Estadística se ocupa del problema de descripción en el caso en que solo es posible observar una fracción -o muestra- de la colección completa de datos que el fenómeno de interés puede producir (habitualmente denominada la población).

En general, las descripciones que produce la Estadística (a nivel exploratorio) se llevan a cabo a través del cálculo de resúmenes de la información disponible. Cuando se trata de un problema de inferencia, la descripción que se obtiene siempre es aproximada puesto que se basa solo en una parte de toda la información que podrá, al menos potencialmente, ser utilizada. En esas condiciones, hay dos retos que es necesario enfrentar. En primer lugar, idealmente, la muestra seleccionada deberá reproducir las características de la población entera. En los términos habituales en la literatura, la muestra deberá ser representativa. En la práctica, sin embargo, es complicado comprobar la representatividad de una muestra ya que ello implicará el conocimiento de la población completa,todo dependerá de si se tiene un buen marco muestral. Por tal razón, en el mejor de los casos, se cuenta con muestras que aproximan el comportamiento de la población y conducen, como ya se indicó, a descripciones aproximadas. El segundo reto consiste precisamente en proveer una medida del grado de aproximación que tienen las inferencias. El trabajo de un puñado de brillantes académicos, entre los que destacan Karl Pearson (1857-1936), Ronald A. Fisher (1890-1962), Egon Pearson (1895-1980), Jerzy Neyman (1894-1981), Harald Cramer (1893-1985), David Blackwell (1919- 2010) y Calyampudi R. Rao (1920-) hizo posible que a lo largo de un periodo de aproximadamente 30 años que inicio alrededor de 1915, los métodos de la Estadística fuesen encontrando respaldo en los principios matemáticos. Es entonces cuando propiamente nace la Estadística Matemática.

Sin embargo, este notable avance de matematización que consolido la Estadística Frecuentista, no resultó en una Teoría, en el sentido axiomático del término. Como sí ocurrió, en cambio, con la Probabilidad en 1933 cuando Andrei Kol-

mogorov (1903-1987) postuló un conjunto de axiomas o principios básicos que encapsulan la naturaleza de la disciplina en su totalidad y a partir de los cuales se pueden deducir todos sus resultados organizados en un cuerpo coherente de conocimientos sin contradicciones ni paradojas.

El surgimiento de una Teoría Estadística o una Teoría de la Inferencia Estadística, habrá de aguardar un tiempo más, hasta la década de los años 50 cuando aparece el libro *The Foundations of Statistics* de Leonard J. Savage (1917-1971) que recoge el fruto de su propio trabajo y el de otros estadísticos como Frank Ramsey (1903-1930), Bruno de Finetti (1906-1985) y Dennis V. Lindley (1923-). En otras palabras, se construye una Teoría Axiomática de la Inferencia Estadística. Probablemente, la consecuencia más importante de este esfuerzo fue el hecho de que la teoría desarrollada, si bien incluye, como casos particulares, algunas ideas, ciertos conceptos y determinados resultados específicos de la poderosa Estadística Frecuentista, en su gran mayoría esta disciplina solo tiene cabida en el nuevo marco como un caso límite.

Así, la nueva teoría nació en conflicto con la escuela predominante de pensamiento estadístico. Más aún, retomó y revaloró ideas y conceptos que habían evolucionado desde finales del siglo XVIII y hasta principios del siglo XX para describir la naturaleza de los fenómenos inciertos. El exponente más brillante de ese enfoque, fue Pierre Simon de Laplace (1749-1827), quien le dio su nombre: Probabilidad Inversa. Laplace elaboró durante años sobre el tema, fue su principal promotor y, en particular, discutió con detalle sus ideas al respecto en obras como: *Memoire sur la probabilité des causes par les événements* de 1774. En algún momento, Laplace dio crédito a un autor que le antecedió en el tratamiento del tema, aunque fuera muy puntual y sin gran repercusión en su tiempo. Ese autor no vió publicado su trabajo ya que este apareció en forma póstuma; su nombre era Thomas Bayes (1702-1761).

Recientemente se ha dado por llamar Neo Bayesiana a la Teoría originada por Ramsey, De Finneti, Lindley y Savage que ha tenido un crecimiento espectacular, especialmente a partir de los ochenta. En una primera fase, la investigación en la materia se orientó al renacimiento de los fundamentos; posteriormente, al desarrollo de métodos Bayesianos para la aplicación en la práctica. Fue esta segunda etapa en la que comprobó que la fortaleza metodológica con frecuencia tenía asociada el costo de la dificultad para obtener resultados con expresiones analíticas cerradas. La tercera etapa, que inicio en los 90, se ha caracterizado por un crecimiento explosivo de las aplicaciones complejas en las más diversas áreas, gracias a la incorporación de técnicas de aproximación numérica vía simulación, especialmente a través de cadenas de Markov.

El propósito de esta tesis es presentar una versión simple pero actualizada de los resultados de las dos primeras etapas y una revisión general de las ideas que gobiernan el desarrollo de la tercera fase. El énfasis se concentra en el procedimiento de construcción de esta Teoría de la Inferencia Estadística (ahora conocida como Bayesiana) así como en ilustrar las principales implicaciones generales que tiene en la práctica y en éste caso es de nuestro interés analizar con esta perspectiva el modelo de regresión lineal.

2. Modelos estadísticos y razonamiento Bayesiano

El propósito principal de la teoría estadística es derivar a partir de observaciones de un fenómeno aleatorio una *inferencia* sobre la distribución de probabilidad subyacente a este fenómeno. Es decir, se proporciona o bien un análisis de un fenómeno pasado, o algunas predicciones sobre un fenómeno futuro de naturaleza similar. Cabe mencionar , en primer lugar, que estos análisis y predicciones suelen ser motivados por un propósito objetivo (si una empresa debe poner en marcha un nuevo producto, un nuevo medicamento debe ser puesto en el mercado, la persona debe vender acciones, etc.) que tienen consecuencias medibles (resultados monetarios, la tasa de recuperación de pacientes, beneficios, etc.). En segundo lugar , proponer procedimientos inferenciales implica tener argumentos sólidos de que son una mejor alternativa que algún otro método.

Insistimos en el hecho de que el proceso de inferencia estadística se debe considerar como una interpretación de los fenómenos, en lugar de una explicación. De hecho, la inferencia estadística se basa en un modelo probabilistico del fenómeno observado e implica un paso de formalización sin el cual no se podría proporcionar una conclusión útil. Es importante mencionar que sólo en este trabajo solo consideraremos modelos paramétricos, suponemos además que las observaciones que apoyan el análisis estadístico, $x_1, x_2, ..., x_n$ han sido generadas por una distribución de probabilidad , i.e., $X_i (1 \leq i \leq n)$ tiene una distribución con densidad $f_i(x_i|\theta_i; x_1, ..., x_{i-1})$ sobre $\mathbb{R}^p$, tal que el parámetro θ_i es desconocido y la función f_i es conocida.

Este modelo puede ser más fácilmente representado por:

$$X \sim f(x|\theta)$$

De esta manera podemos decir que un modelo estadístico paramétrico consiste de la observación de una variable aleatoria X , distribuida de acuerdo a $f(x|\theta)$, donde solo el parámetro θ es desconocido y pertenece a un espacio vectorial Θ de dimensión finita.

Una vez definido el modelo estadístico, el propósito principal del análisis estadístico consiste en llevar a una inferencia sobre el parámetro θ. Esto significa que utilizamos la observación x para mejorar nuestro conocimiento sobre el parámetro θ, de modo que uno puede tomar una decisión relacionada con este parámetro. Este conocimiento corresponde a estimar una función de θ o un evento futuro que depende de la distribución θ.

En términos más generales, la inferencia cubre el fenómeno aleatorio dirigido por θ y por lo tanto incluye la predicción, es decir, la evaluación de la distribución de una futura observación y dependiendo de θ(y posiblemente la observación actual x), $y \sim g(y|\theta, x)$.

La elección Bayesiana puede aparecer como una reducción innecesaria del alcance inferencial, y de hecho ha sido criticada por muchos. Pero veremos a lo largo de los siguientes capítulos que esta reducción es a la vez necesaria y beneficiosa.[16]

2.1. El paradigma Bayesiano

En comparación con el modelado probabilístico, el propósito de un análisis Bayesiano es fundamentalmente un propósito de inversión, ya que tiene como objetivo la recuperación de la causas a través de las observaciones.

En otras palabras, al observar un fenómeno aleatorio dirigido por un parámetro θ, los métodos estadísticos permiten deducir de estas observaciones una inferencia (que es, un resumen, una caracterización) sobre θ. Este aspecto de inversión de la estadística es claro en la noción de función de verosimilitud, puesto que, formalmente, es sólo la densidad de la muestra reescrita como función de θ.

$$L(\theta|x) = f(x|\theta)$$

2.1.1. Teorema de Bayes

El teorema de Bayes es el puente para pasar de una probabilidad a priori o inicial de una hipótesis a una probabilidad a posteriori o actualizada, basado en una nueva observación . 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 descripción general de la inversión de probabilidades esta dada por el teorema de Bayes que a continuación se enuncia:

Teorema 2.1. ***De Bayes***
Sean A y B eventos tal que $P(B) \neq 0$, entonces $P(A|B)$ y $P(B|A)$ están relacionadas por :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{P(B|A)P(A)}{P(B)}$$

En particular,

$$\frac{P(A|B)}{P(C|B)} = \frac{P(B|A)}{P(B|C)}$$

cuando $P(A) = P(C)$.

Demostración:
El teorema de Bayes, se deriva de manera casi inmediata a partir de la definición de probabilidad condicional; consideremos dos eventos A y B con probabilidad

no nula, entonces la probabilidad condicional de A dado B, se define del modo siguiente:

$$P\left(A|B\right) = \frac{P\left(A\bigcap B\right)}{P\left(B\right)},\ con\ P\left(B\right)\neq 0 \tag{2.1}$$

Análogamente,

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

Así que sustituyendo en 2.1 la expresión $P\left(A\bigcap B\right)=P\left(B|A\right)P\left(A\right)$, tenemos que

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}\blacksquare$$

La demostración del Teorema de Bayes es sencilla usando la axiomática moderna de la probabilidad, sin embargo, este aparece como uno de las principales pasos conceptuales en la historia de la estadística, siendo la primera inversión de probabilidades.

Ahora supongamos que $A_1, A_2, ..., A_k$ son k eventos mutuamente excluyentes, uno de los cuales ha de ocurrir necesariamente, entonces la ley de la probabilidad total establece que:

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

Si consideramos el evento A_j en lugar de A en el Teorema de Bayes y aplicando al denominador la ley de la probabilidad total tenemos

$$P(A_j|B) = \frac{P\left(B|A_j\right)P\left(A_j\right)}{\sum_{i=1}^{k} P\left(B|A_i\right)P\left(A_i\right)},\ j=1,...k \tag{2.2}$$

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

Se tiene una versión continua de este resultado, sean Y y X variables aleatorias con densidad condicional $f(x|y)$ y densidad marginal $g(y)$, la densidad condicional de Y dado X esta dada por:

$$g\left(y|x\right) \quad = \quad \frac{f(x|y)g(y)}{\int f(x|y)g(y)dy}$$

Aunque este teorema de inversión es muy natural desde el punto de vista probabilista, Bayes y Laplace fueron más allá y consideran que la incertidumbre sobre el parámetro θ de un modelo podría ser descrita a través de una distribución de probabilidad π sobre Θ , llamada distribución *a priori*. La inferencia se basa entonces en la distribución de θ condicionada a x , $\pi(\theta|x)$, llamada distribución *a posteriori* definida como:

$$\pi(\theta|x) \quad = \quad \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Note que $\pi(\theta|x)$ es proporcional a la distribución de x condicionada a θ, i.e., la verosimilitud multiplicada por la distribución a priori.

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$

La principal contribución de un modelo estadístico Bayesiano es considerar una distribución de probabilidad para los parámetros.

Formalmente podemos decir que un modelo estadístico Bayesiano parámetrico esta formado por un modelo estadístico paramétrico, $f(x|\theta)$, y la distribución *a priori* sobre los parámetros,$\pi(\theta)$.

En términos estadísticos, el Teorema de Bayes actualiza la información sobre θ, mediante la extracción de la información contenida en θ en la observación x. Su impacto se basa en el movimiento audaz que pone causas (observaciones) y efectos (parámetros) en el mismo nivel conceptual, ya que ambos tienen distribuciones de probabilidad. Desde un punto de vista de modelado estadístico, hay por lo tanto poca diferencia entre las observaciones y parámetros, ya que al condicionar se tiene una interacción de sus respectivas distribuciones.

2.1.2. Intercambiabilidad

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. Este concepto permite, desde una perspectiva subjetiva, justificar el uso (así como mejorar la interpretación) de algunos conceptos estadísticos comunes, tales como parámetro, muestra aleatoria, verosimilitud y distribución inicial.

El supuesto de intercambiabilidad, a grandes rasgos significa que los n valores x_i observados en la muestra pueden ser intercambiados, es decir, que la distribución

conjunta $P(X_1, ..., X_n)$ debe ser invariante a las permutaciones de los índices. Generalmente, es útil modelar los datos de una distribución intercambiable como independientes e idénticamente distribuidas dado algún vector de parámetros desconocidos θ con distribución $\pi(\theta)$.

Definición 2.1. *Sea $X_1, ..., X_n$, una sucesión de variables aleatorias. Diremos que son intercambiables bajo una medida de probabilidad P, si para cada subconjunto finito de tal sucesión, su distribución conjunta satisface:*

$$P(X_1, ..., X_n) = P\left(X_{\tau(1)}, ..., X_{\tau(n)}\right)$$

para todas las permutaciones τ definidas sobre el conjunto $\{1, ..., n\}$

La definición anterior establece que toda la información relevante, en un conjunto de variables intercambiables, está contenida en los valores de las $x_i's$, 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.

Utilizando el concepto de intercambiabilidad, de Finetti demuestra su teorema de representación para variables dicotómicas:

Teorema 2.2. ***Teorema de representación para variables dicotómicas*** *Si $\{X_1, ..., X_n\}$ son variables aleatorias intercambiables y $x_i \in \{0, 1\}$ $\forall i = 1, ..., n$, entonces:*

1. *Existe $\theta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i$, $\theta \in (0, 1)$*
2. *Existe una densidad de probabilidad $\pi(\theta)$, tal que la densidad conjunta $P(X_1, ..., X_n)$ tiene la representación integral*

$$P(X_1, ..., X_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \pi(\theta) \, d\theta$$

El teorema muestra que 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 cual existe una distribución inicial $\pi(\theta)$.

En el caso dicotómico, la intercambiabilidad identifica las observaciones como una muestra aleatoria de un modelo probabilístico especifico y garantiza la existencia de una distribución inicial sobre su parámetro. En el caso general, para variables aleatorias de cualquier rango y dimensión, la intercambiabilidad identifica las observaciones como una muestra aleatoria de algún modelo probabilístico y garantiza la existencia de una distribución inicial sobre el parámetro que lo describe.

Teorema 2.3. ***Teorema de representación general***
Si $\{X_1, ..., X_n\}$ *son variables aleatorias intercambiables y* $x_i \in X$*, entonces:*

1. *Existen una función* $f(x_1, ..., x_n)$ *y un modelo probabilístico* $P(x|\theta)$*, con* $x \in X$ *, donde* $\theta = \lim_{n\to\infty} f(x_1, ..., x_n)$*,* $\theta \in \Theta$

2. *Existe una función de densidad de probabilidad* $P(\theta)$ *tales que la densidad conjunta* $P(X_1, ..., X_n)$ *tiene la representación integral*

$$P(X_1, ..., X_n) = \int_\Theta \prod_{i=1}^{n} P(X_i|\theta) P(\theta) \, d\theta$$

En una terminología convencional, el teorema demuestra que un conjunto de observaciones intercambiables pueden ser siempre interpretadas como una muestra aleatoria de algún modelo probabilístico, controlado por un parámetro que no es introducido de manera arbitraria, sino que resulta definido como el límite de una función de las observaciones. Además, el teorema garantiza la existencia de una distribución inicial sobre el parámetro que gobierna el comportamiento del modelo. Debe subrayarse que las observaciones intercambiables x_i pueden ser vectores de cualquier dimensión y que, incorporando en ellas como covariables todos los elementos de información relevantes para que sus índices no sean informativos, siempre es posible postular la intercambiabilidad de los vectores así obtenidos.

Los teoremas de representación son resultados de teoría de la probabilidad no sujetos a polémica alguna, pero al demostrar la existencia de la distribución inicial constituyen una demostración de la necesidad lógica de la metodología Bayesiana. Observe, sin embargo, que se trata tan sólo de teoremas de existencia: en cada problema concreto es necesario identificar tanto el modelo como la distribución inicial apropiadas.

2.2. Principios de suficiencia y máxima verosimilitud

2.2.1. Suficiencia

Los estadísticos clásicos la mayoría de las veces suelen basarse en el sentido común o axiomas adicionales; por el contrario, el enfoque Bayesiano incorpora naturalmente la mayoría de estos principios sin la restricción sobre los procedimientos y también rechaza definitivamente otros principios, como el de insesgamiento. Esta noción fue una vez uno de los pilares de la estadística clásica y limita la elección de los estimadores para los que son en promedio correctos. Mientras que, intuitivamente aceptable, impone condiciones muy estrictas sobre la elección de los procedimientos y con frecuencia conduce a la ineficiencia en sus actuaciones. Más importante aún, el número de problemas que permiten soluciones insesgadas es un porcentaje insignificante de todos los problemas de estimación.

Dos principios fundamentales son seguidos por el paradigma Bayesiano, a saber, el Principio de Verosimilitud y el Principio de Suficiencia.

Definición 2.2. *Sea $X \sim f(X|\theta)$, una función T de X (llamada estadística) se dice que es suficiente si la distribución de X condicionada a $T(X)$ no depende de θ.*

Una estadística suficiente $T(X)$ contiene toda la información presentada por X acerca de θ. Según el teorema de factorización, bajo condiciones de regularidad, la densidad de X puede entonces ser escrita como:

$$f(X|\theta) = g(T(X)|\theta)h(X|T(X))$$

si g es la densidad de $T(X)$.

El concepto de suficiencia ha sido desarrollado por Fisher y se asocia con el siguiente principio.

Principio de Suficiencia

Dos observaciones x y y que mediante una estadística suficiente T llevan al mismo valor (esto es $T(x) = T(y)$), deben conducir a la misma inferencia sobre θ.

El Principio de Suficiencia es generalmente aceptado por la mayoría de los estadísticos, en particular debido al teorema de Rao-Blackwell, que descarta estimadores que no dependen sólo de estadísticas suficientes. En la elección del modelo , a veces es criticado por ser demasiado reductiva, pero tengamos en cuenta que el Principio de Suficiencia sólo es legítimo cuando el modelo estadístico es en realidad el que subyace a la generación de las observaciones.

Cualquier incertidumbre sobre la distribución de las observaciones debe ser incorporada en el modelo, una modificación es casi seguro que conduce a un cambio de estadísticas suficientes. Una observación de advertencia similar se aplica al Principio de verosimilitud.

2.2.2. Principio de máxima verosimilitud

Este segundo principio es en parte consecuencia del principio de suficiencia. Este se puede atribuir a Fisher (1959) o incluso a Barnard (1949), pero fue formalizado por Birnbaum (1962). Es fuertemente defendido por Berger y Wolpert (1988) quienes proporcionan un estudio extenso del tema. En la siguiente definición , la noción de información debe ser considerada en el sentido general de la colección de todas las deducciones posibles de θ, y no en el sentido matemático de la información de Fisher definida más adelante.

Principio de verosimilitud

La información presentada por una observación x acerca de θ está contenida enteramente en la función de verosimilitud $L(\theta|x)$. Por otra parte , si x_1 y x_2

son dos observaciones en función del mismo parámetro θ , tal que existe una constante c y se tiene que

$$L(\theta|x_1) \quad = \quad c\, L(\theta|x_2)\,,$$

para toda θ, es decir, tener la misma información sobre θ debe dar lugar a inferencias idénticas.[16]

Observe que el Principio de Verosimilitud es válido cuando:

- la inferencia es sobre el mismo parámetro θ,
- θ incluye todos los factores desconocidos del modelo.

2.2.3. Derivación del principio de verosimilitud

Una justificación del principio de verosimilitud ha sido proporcionada por Birnbaum (1962), que estableció qué se da a entender por el Principio de Suficiencia, condicional a la aceptación de un segundo principio.

Principio de condicionalidad

Si dos experimentos sobre el parámetro θ, ε_1y ε_2, están disponibles y si se selecciona uno de estos dos experimentos con probabilidad p, la inferencia resultante sobre θ sólo debe depender del experimento seleccionado. Este principio parece difícil de rechazar cuando el experimento seleccionado es conocido.

Teorema 2.4. *El Principio de Verosimilitud es equivalente a la conjunción de los Principios de Condicionalidad y Suficiencia.*

2.3. Distribuciones a priori y a posteriori

2.3.1. Distribución a priori o inicial

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 determinara 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 del análisis Bayesiano, sin embargo son el punto más criticado de este. La elección de la distribución inicial es lo más importante para hacer inferencia y hasta cierto punto es lo más complicado. En comparación con la inferencia clásica en donde se tiene una muestra aleatoria con parámetros fijos, en el enfoque Bayesiano los parámetros son aleatorios, lo cuál es simplificado en una distribución inicial.

2.3.2. Priori no informativa

Cuando nada es conocido sobre los parámetros, la selección de una distribución a priori adecuada adquiere una connotación especial pues será necesario elegir una distribución a priori que no influya sobre ninguno de los posibles valores de los parámetros en cuestión. Estas distribuciones a priori reciben el nombre de difusas o no informativas y en esta sección se tratara algunos criterios para su selección.

De manera más formal podemos decir que 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, la distribución a priori $\pi(\theta)$ es una distribución uniforme, es decir, todos los posibles resultados de θ tienen la misma probabilidad.

El caso más simple se presenta cuando el espacio parametral Θ es un conjunto finito. Si Θ consiste de n elementos , entonces la distribución inicial no informativa más simple es la que asigna a cada elemento de Θ la probabilidad de $\frac{1}{n}$.

Esta idea puede generalizarse, supongamos que Θ es un conjunto infinito, se puede asignar a cada $\theta \in \Theta$ la misma probabilidad, obteniendo así, la distribución a priori no informativa uniforme $\pi(\theta) = c$.

El principal problema de emplear la distribución uniforme como a priori no informativa , es que la distribución uniforme es no invariante bajo reparametrizaciones, es decir, si no se cuenta con información sobre θ , entonces tampoco sobre $g(\theta)$ así:

$$\text{Si } \pi(\theta) = c \text{ y } \phi = g(\phi) \Longrightarrow \pi(\theta) = \det\left(\tfrac{d}{d\theta} g^{-1}(\phi)\right) \pi\left(g^{-1}(\phi)\right)$$

También, cuando el espacio parámetrico Θ es infinito, la distribución uniforme presenta el problema de ser impropia, aunque no es un problema serio pues frecuentemente una a priori impropia lleva a una posterior propia.

2.3.3. Método de Jeffreys

En situaciones generales, para un parámetro θ , el método más usado es el de Jeffreys, este sugiere que si un investigador no conoce información acerca de θ, entonces lo que espera observar de θ, dada la información muestral Y debe

ser la misma que la de una parametrización para θ o cualquier tranformación inyectiva de θ, $g(\theta)$. Una a priori invariante sería:

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

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 $\theta = (\theta_1, ..., \theta_n)'$ entonces:

$$\pi(\theta) \propto \sqrt{\det(I(\theta))}$$

con $I(\theta)$ la matriz de información de Fisher de orden $p \times p$, y la entrada (i, j) es:[6]

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

2.3.4. Distribución a priori conjugada

En este caso, la distribución a priori es determinada completamente por una función de densidad conocida. Se ha dicho que el conocimiento sobre θ puede emplearse para especificar una distribución inicial con una forma particular. Así, una familia parámetrica 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 presenta la siguiente definición para una familia conjugada:

Definición 2.3. *Una clase $\mathcal{L}$ de distribuciones a priori es denominada una familia conjugada para la clase de funciones de densidad $\mathcal{F}$, si $f(\theta|y)$ está en en la clase $\mathcal{L}$ para todo $f(y|\theta) \in \mathcal{F}$ y $\pi(\theta) \in \mathcal{L}$.*

Definición 2.4. *Sea $Y_1, ..., Y_n$ una muestra aleatoria de $f(Y|\theta)$, una familia* T *de densidades se dice que es conjugada para la función de densidad $f(Y|\theta)$ si la función de densidad a priori de θ , $\pi(\theta) \in$ T y $f(\theta|y) \in$ T.*

En este caso, la distribución inicial dominará a la función de verosimilitud y $f(\theta|y)$ tendrá la misma forma que $\pi(\theta)$, con los parámetros corregidos por la información muestral. Generalmente para una clase de densidades $\mathcal{F}$, una familia conjugada puede ser determinada examinando la 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.

2.3.5. Distribución a posteriori

La información a posteriori de θ dado y , con $\theta \in \Theta$, está dada por la expresión $\pi(\theta|y)$ y expresa lo que es conocido de θ después de observar los datos de y. Por la ley multiplicativa de la probabilidad θ y y tienen la siguiente función de densidad subjetiva conjunta

$$h(y,\theta) = \pi(\theta) L(y|\theta)$$

en donde:

$\pi(\theta)$: densidad a priori de θ,

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

y tienen función de densidad marginal dada por:

$$m(y) = \int_{\Theta} \pi(\theta) L(y|\theta) \, d\theta$$

Si la función marginal $m(y) \neq 0$

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

Sustituyendo

$$\pi(\theta|y) = \frac{\pi(\theta) L(y|\theta)}{\int_{\Theta} \pi(\theta) L(y|\theta) \, d\theta} \tag{2.3}$$

La cuál resulta ser una de las versiones del Teorema de Bayes.

Dada una muestra de n observaciones independientes se puede obtener $L(y|\theta)$ y evaluar en la ecuación 2.3. La evaluación de esta expresión puede ser simplificada , utilizando un estadístico suficiente para θ con función de densidad $g(S(y)|\theta)$, lo cual se enuncia en el lema siguiente:

Lema 2.1. *Sea $S(y)$ un estadístico suficiente para θ (por lo que $L(y|\theta) = h(y)\, g(S(y)|\theta)$) y $m(s) \neq 0$ la densidad marginal para $S(y) = s$), entonces:*

$$\pi(\theta|y) \quad = \quad \pi(\theta|s) = \frac{g(s|\theta)\,\pi(\theta)}{m(s)}$$

Demostración:
Tenemos que

$$\begin{aligned}
\pi(\theta|y) &= \frac{\pi(\theta)\, L(y|\theta)}{\int_{\Theta} \pi(\theta)\, L(y|\theta)\, d\theta} \\
&= \frac{\pi(\theta)\, h(y)\, g(S(y)|\theta)}{\int_{\Theta} \pi(\theta)\, h(y)\, g(S(y)|\theta)\, d\theta} \\
&= \frac{g(s|\theta)\,\pi(\theta)}{m(s)} = \pi(\theta|s)\, \blacksquare
\end{aligned}$$

Podemos observar que la ecuación 2.3 se puede expresar convenientemente como:

$$\pi(\theta|y) \quad \propto \quad L(y|\theta)\,\pi(\theta) \tag{2.4}$$

En la práctica , el cálculo de la distribución a posteriori puede resultar una tarea complicada, especialmente si la dimensión del vector de parámetros es grande. Sin embargo, para ciertas distribuciones iniciales y verosimilitudes es posible simplificar el análisis. En otros casos se requieren aproximaciones analíticas o el uso de técnicas de simulación.

3. Inferencia Bayesiana

Dado que la distribución posterior, contiene toda la información concerniente al parámetro de interés θ (información a priori y muestral), cualquier inferencia con respecto a θ consistirá en afirmaciones hechas a partir de dicha distribución. En éste capítulo se expondrán los diferentes tipos de estimación necesarios para el análisis Bayesiano.

3.1. Estimación por máxima verosimilitud

Para estimar el parámetro θ, se pueden aplicar técnicas de la estadística clásica a la distribución a posteriori. La técnica más utilizada 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 $L(\theta|y)$. Análogamente, la estimación Bayesiana por máxima verosimilitud se define de la siguiente manera:

Definición 3.1. *El estimador generalizado de máxima verosimilitud de θ es definido como la moda más grande , $\hat{\theta}$ de la distribución a posteriori $\pi(\theta|y)$.*

3.1.1. Error de estimación

Cuando se hace una estimación , usualmente necesitamos indicar la precisión de la estimación. En el caso univariado la medida Bayesiana que se utiliza para medir la precisión de una estimación es la varianza a posteriori del estimador.

Definición 3.2. *Si θes un parámetro de valor real con distribución a posteriori $\pi(\theta|y)$ y δ es un estimador de θ, entonces la varianza a posteriori de δ es :*

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

Notemos que en la definición anterior, al tener el estimador $\hat{\theta}$ de θ, solo se necesita sustituir en la definición de la varianza para obtener una expresión de 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 nos serán útiles más adelante , se puede representar a la varianza a posteriori de la siguiente forma:

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

De esta manera

$$V_{\delta}^{\pi}(y) = V^{\pi}(y)+(\mu^{\pi}(y)-\delta)^2 \tag{3.1}$$

3.1.2. Estimación multivariada

La estimación Bayesiana de un vector $\theta=(\theta_1,...,\theta_n)$ es la estimación de máxima verosimilitud generalizada (moda a posteriori) , $\hat{\theta}=\left(\hat{\theta}_1,...,\hat{\theta}_n\right)$. Sin embargo, no podemos garantizar la existencia y unicidad en el caso multivariado, por ello es más conveniente utilizar la medida a posteriori

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

como el estimador Bayesiano,y la precisión puede ser descrita por la matriz de covarianzas a posteriori

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

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

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

3.2. Estimación puntual

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 a posteriori. Así dado un conjunto de datos Y , todo lo que podemos decir sobre alguna función de θ, que rige el modelo, estará contenida en la distribución posterior $\pi(\theta|Y)$.

La inferencia Bayesiana puede ser descrita técnicamente como un problema de decisión donde el espacio de acciones es la clase de las distribuciones a posteriori del parámetro de interés que son compatibles con las suposiciones aceptadas.

La distribución posterior reemplaza la función de verosimilitud como una expresión que incorpora toda la información. $f(\theta|y)$ 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 estimados puntuales de los parámetros son la consecuencia natural de una inferencia).

En el contexto Bayesiano se puede reducir la información de $f(\theta|y)$ a un simple "mejor" estimado; existen dos formas de enfrentar el problema:

- Estimador de Bayes posterior
- Teoría de la decisión

3.2.1. Estimador de Bayes posterior

El estimador de Bayes posterior se define de la siguiente manera:

Definición 3.3. *Sea* $\{X_1, X_2, ..., X_n\}$*una muestra aleatoria de* $f(X_i|\theta)$*, donde* θ *es un valor de la variable aleatoria* Θ *con densidad* $\pi(.)$*. El estimador de Bayes posterior de* $\tau(\theta)$ *con respecto a la priori* $\pi(.)$ *es definida como*

$$E(\tau(\theta)|x_1, x_2, ..., x_n) = \frac{\int_\Theta \tau(\theta) \left[\prod_{i=1}^{n} f(x_i|\theta)\right] \pi(\theta) d\theta}{\int_\Theta \left[\prod_{i=1}^{n} f(x_i|\theta)\right] \pi(\theta) d\theta}$$

Una vez que hemos calculado $E(\tau(\theta)|x_1, x_2, ..., x_n)$, encontramos el estimador Bayesiano para θ, minimizando alguna función de pérdida. Esto es:

$$\hat{\theta} = argminE(L(\theta, \delta(x_1, ..., x_n)))$$

3.2.2. Aproximación vía teoría de decisión

La teoría de la 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 más elevado (esto puede ser interpretado como elegir la decisión más acertada).[1]

La idea es combinar la información muestral con información no muestral con el objetivo de tomar la decisión óptima. El análisis Bayesiano comparte con la teoría de la decisión el uso de información no muestral.

Sean Θ el espacio parametral y A el espacio de acciones , y entendamos que $a \in A$ es una acción. Un problema de decisión es una función $D: \ A \times \Theta \to C$, donde C es un espacio de consecuencias. Las preferencias sobre el conjunto de consecuencias las representaremos mediante una función de pérdida.

Definición 3.4. *Sea $L(\theta, a)$ una función de pérdida.*
La pérdida esperada Bayesiana cuando se toma una decisión $a \in A$ es:

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

Y el estimador Bayesiano se define como:

$$\hat{\theta} \quad = \quad argmin \ p(a|y)$$

Funciones de pérdida

Se especifica una función de pérdida $L(\theta, a)$ que cuantifica las posibles penalidades en estimar θ por medio de a. Hay muchas funciones de pérdida que se pueden utilizar , la elección de alguna de ellas dependerá del contexto del problema.

Las más utilizadas son:

- Pérdida cuadrática: $L(\theta, a) = (\theta - a)^2$
- Pérdida de error absoluto: $L(\theta, a) = |\theta - a|$
- Pérdida 0,1: $L(\theta, a) = \begin{cases} 1 & si \ |\theta - a| \leq \varepsilon \\ 0 & si \ |\theta - a| > \varepsilon \end{cases}$
- Pérdida lineal: $L(\theta, a) = \begin{cases} g(a - \theta) & si \ a > \theta \\ g(\theta - a) & si \ a < \theta \end{cases}$

Para 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 estimador puntual de θ para la elección en particular de la función de pérdida.

Ejemplo 3.1. *Sea* $\{X_i\}_{i=1}^n$ *una muestra aleatoria de la distribución* $Bin(\theta, 1)$. *Supongamos que* $\pi(\theta) \sim Beta(a, b)$ *es decir*

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} I_{(0,1)}(\theta)$$

Sabemos que

$$\begin{aligned} E(\theta) &= \frac{a}{a+b} \\ Var(\theta) &= \frac{ab}{(a+b)^2(a+b+1)} = \frac{E(\theta)(1-E(\theta))}{a+b+1} \end{aligned} \tag{3.2}$$

La expresión 3.2 muestra que para un valor fijo de $E(\theta)$ *la varianza depende de* $a+b$ *y* $Var(\theta) \to 0$ *cuando* $a+b \to \infty$.
Por otro lado la función de verosimilitud está dada por:

$$L(\underline{X}|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

De la verosimilitud y la distribución a priori podemos inferir que:

$$\pi(\theta|\underline{x}) = \frac{\theta^{\sum_{i=1}^n x_i + a - 1}(1-\theta)^{n-\sum_{i=1}^n x_i + b - 1}}{\int_0^1 t^{\sum_{i=1}^n x_i + a - 1}(1-t)^{n-\sum_{i=1}^n x_i + b - 1}\, dt}$$

observe que $\pi(\theta|x)$ *difiere por alguna contante que es función de la muestra de una distribución* $Beta(a + \sum_{i=1}^n x_i, n - \sum_{i=1}^n x_i + b)$. *Por tanto podemos afirmar que*

$$\pi(\theta|x) \sim Beta\left(a + \sum_{i=1}^n x_i, n - \sum_{i=1}^n x_i + b\right)$$

Luego el estimador de Bayes , que denotaremos por $\delta_{a,b}$ esta dado por $E(\theta|x)$ y tenemos que:

$$\delta_{a,b} = \frac{T+a}{a+b+n} = \frac{n}{n+a+b}\frac{T}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b}$$

donde $T = \sum_{i=1}^{n} x_i$. Notemos que el estimador de Bayes se comporta como un promedio ponderado de dos estimadores $\delta_1 = \frac{T}{n}$ que no utiliza la información de la distribución a priori y $\delta_2 = \frac{a}{a+b}$ que corresponde a la esperanza de la distribución a priori . El peso asignado a δ_2 tiende a cero cuando $n \to \infty$

3.3. Intervalos de credibilidad

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 estimadores puntuales no proporcionan una medida de la precisión de la estimación. Esto genera problemas en el enfoque clásico desde que los parámetros no son considerados como aleatorios, por lo que no es posible dar un intervalo con la interpretación de que existe una cierta probabilidad de que el parámetro este en el intervalo. Por otro lado en el enfoque Bayesiano , no se tiene esa dificultad pues los parámetros son aleatorios.

Definición 3.5. *Un conjunto veraz al $100(1-\alpha)$ % para θ es un subconjunto C de Θ tal que:*

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

Al igual que sucede con los intervalos de confianza en el enfoque clásico , los conjuntos veraces no son únicos.

La distribución a posteriori es una distribución de probabilidad en Θ, e intuitivamente podemos decir que la probabilidad de θ esta 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 creíble $C(Y)$ contenga a θ.

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

Para lograr lo antes descrito, solo debemos de considerar aquellos puntos con los valores de $f(\theta|y)$ más grandes. Esto conduce a un intervalo de la forma:

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

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

La región C que cumple las anteriores condiciones se denomina "región de densidad posterior más grande"y por sus siglas en inglés lo llamaremos HPD.

Generalmente un HPD es encontrado con la ayuda de métodos numéricos y la dificultad depende de la dimensión del vector de parámetros.

3.3.1. Contraste entre el enfoque clásico y el Bayesiano

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 , el conjunto C contenga a θ con al menos una probabilidad de $(1-\alpha)\,100\,\%$.

Por otro lado, un intervalo de confianza al $(1-\alpha)\,100\,\%$ del enfoque clásico 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 muestras 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 se pueden generar tantas muestras aleatorias en la práctica.

3.3.2. Intervalo de mínima longitud-máxima densidad

Para poder encontrar el intervalo de mínima longitud se debe de:

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 $f(\theta|y)$, es frecuentemente conveniente dar regiones $R \subset \Theta$ de probabilidad dadas en virtud de $f(\theta|y)$. Por ejemplo , la identificación de regiones que contienen $50\,\%$, $90\,\%$, $95\,\%$ o $99\,\%$ de probabilidad.

Una región $R \subset \Theta$ tal que

$$\int_R f(\theta|y) \quad = \quad q$$

(dado y , el valor verdadero de R con probabilidad q), se dirá que es una región $q-$*créible* posterior de θ.

Note que que esta noción proporciona inmediatamente una afirmación directa e intuitiva sobre la cantidad desconocida de interés θ, en términos de probabilidad , en marcado contraste con las afirmaciones previstas con los intervalos de confianza frecuentistas.

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, es decir, la imagen $\phi(R)$ de una región R de HPD , será una región de credibilidad para ϕ , pero generalmente no será HPD. Los cuantiles a posteriori son frecuentemente usados para derivar regiones de credibilidad.

Así $\theta_q = \theta_q(y)$ es el cuantil posterior q de θ, $R = \{\theta|\theta \leq \theta_q\}$ es una región $q - creíble$ unilateral, generalmente única. Más aún, las regiones $q - creíbles$ 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 más empleadas que las regiones HPD.

Análogamente a la estimación puntual, una región de credibilidad Bayesiana se define de la siguiente manera.

Definición 3.6. *Una región $R \subseteq \Theta$ tal que,*

$$\int_R f(\theta|y)\, d\theta \quad = \quad 1-\alpha$$

se dirá que es una $100(1-\alpha)\,\%$ región de credibilidad para θ , con respecto a $f(\theta|y)$.

Es claro que para toda α no existe una única región de credibilidad. Por tanto, el problema de elegir entre los subconjuntos de $R \subseteq \Theta$ tal que $\int_R f(\theta|y)\, d\theta = 1-\alpha$, para algún θ, $f(\theta|y)$ 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 3.1. *Sea $f(\theta|y)$ 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:*

$$\forall \theta_1 \in R, \quad \theta_2 \notin R \quad : \quad f(\theta_1|y) \geq f(\theta_2|y)$$

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

Definición 3.7. *Una región $R \subseteq \Theta$ será llamada una $100\,(1-\alpha)$ % región con función de densidad de probabilidad alta para θ con respecto a $f\,(\theta|y)$ si:*

1. $P\,[\theta \in R] = 1 - \alpha$
2. $\forall \theta_1 \in R, \theta_2 \notin R:\ f\,(\theta_1|y) \geq f\,(\theta_2|y)$ *excepto en un subconjunto de Θ de probabilidad cero.*

3.4. 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. En la estadística clásica para el contraste de hipótesis se tienen, la hipótesis nula $H_0:\ \theta \in \Theta_0$ y la hipótesis alternativa $H_1:\ \theta \in \Theta_1$, donde $\Theta_0$es el espacio parametral bajo la hipótesis nula y Θ_1 el espacio parametral bajo la hipótesis alternativa.

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 dada que es falsa, respectivamente.

En la estadística Bayesiana , la tarea de decidir entre H_0 y H_1 , es más simple , pues 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 distribuciones a priori elegidas. Denotaremos con π_0 y π_1 a las probabilidades a priori de Θ_0 y Θ_1 respectivamente.

Definición 3.8. *Se denomina razón a priori de H_0 contra H_1 al cociente*

$$\frac{\pi_0}{\pi_1}$$

Análogamente, razón a posteriori de H_0contra H_1 al cociente

$$\frac{\alpha_0}{\alpha_1}$$

Si la razón a priori es cercana a 1 , H_0 y H_1 tienen inicialmente casi la misma probabilidad de ocurrir. Si este cociente es mayor a 1 se considera que H_0 es más probable que H_1 y si el cociente es menor a 1 inicialmente se considera que H_0 es menos probable que H_1.

Definición 3.9. *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 a favor de Θ_0.

Nos interesa el factor de Bayes, ya que en algunas ocasiones puede ser interpretado como la razón entre H_0 y H_1 ,dados los datos. Esta interpretación es válida cuando se tienen hipótesis simples, es decir , cuando $\Theta_0 = \{\theta_0\}$ y $\Theta_1 = \{\theta_1\}$. Para este caso α_0 y $\alpha_1$están dadas por :

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

por lo que la razón a posteriori es

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

y el factor de Bayes

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

En otras palabras, B es la razón de probabilidad de H_0 contra H_1. Note 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 H_1. De esta manera, Si $B > 1$ quiere decir que H_0 es ahora relativamente más probable que H_1 y si $B < 1$ es ahora H_1 la que es relativamente más probable.

3.4.1. Contraste unilateral

Este tipo de contraste se da cuando $\Theta \subseteq \mathbb{R}$ y los contrastes son de la siguiente naturaleza:

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

o el otro caso

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

En las pruebas de hipótesis del enfoque clásico se puede calcular un estadístico que permite resumir el resultado del experimento de manera objetiva. Esta cantidad es el *p-valor* que corresponde al *nivel de significancia más pequeño posible* que puede escogerse, para el cual todavía *no se rechazaría la hipótesis alternativa* con las observaciones actuales.

Lo interesante de este tipo de contraste es que el uso del *p-valor* del enfoque clásico tiene una justificación Bayesiana. Este hecho se ilustra con el siguiente ejemplo.

Ejemplo 3.2. *Sea* $Y \sim N\left(\theta, \sigma^2\right)$ *y supongamos que* θ *tiene una distribución a priori no informativa* $\pi(\theta) \propto 1$*, entonces*

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

El contraste que deseamos realizar es el siguiente:

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

luego

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

el $p - valor$ *está dado por*

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

Ya que la distribución normal es simétrica, se tiene que el $p-valor$ *es el mismo que* α_0*.*

3.4.2. Contraste de hipótesis nula puntual

El contraste de hipótesis nula puntual se refiere al siguiente contraste:

$$H_0 : \ \theta = \theta_0 \quad vs \quad H_1 : \ \theta \neq \theta_0$$

Este contraste es interesante porque las respuestas en el enfoque Bayesiano difieren radicalmente de las que da el enfoque clásico.

En pocas ocasiones se tiene que $\theta - \theta_0 = 0$. Es más común que se cumpla la hipótesis nula como $\theta \in \Theta_0 = (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, en donde $\varepsilon > 0$ y es tal que para cada $\theta \in \Theta_0$, θ y θ_0 sean indistinguibles.

Lo ideal es elegir ε como un valor real pequeño , de lo contrario θ y θ_0 diferirán cada vez más. Dado que se debe hacer el contraste $H_0 : \ \theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ es importante saber cuando es apropiado que se aproxime de esa forma $H_0 : \ \theta = \theta_0$.

Desde la perspectiva del análisis Bayesiano, la respuesta a esta interrogante es, la aproximación es razonable si las probabilidades a posteriori de H_0 son casi iguales para ambos casos. Una condición fuerte cuando éste sea el caso es que la función de distribución conjunta de probabilidad observada sea aproximadamente constante en $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$.

Para llevar a cabo el contraste Bayesiano para una hipótesis nula puntual , no se utiliza una densidad a priori continua , ya que en cualquiera de estos casos $\theta_0$tendrá una probabilidad a priori igual a cero. Se debe utilizar una distribución inicial mixta que asigne una probabilidad de π_0 al punto θ_0 y el resto se distribuya 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 de la variable Y, la densidad marginal de $Y = (Y_1, ..., Y_n)$ es,

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

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

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

luego

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

Entonces el factor de Bayes de H_0 contra H_1 es

$$\begin{aligned} B &= \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} \\ &= \frac{\pi_0 f\left(y|\theta\right) \pi_1}{\pi_1 m_1\left(y\right) \pi_0} \\ &= \frac{f\left(y|\theta_0\right)}{m_1\left(y\right)} \end{aligned}$$

$$B = \frac{f\left(y|\theta_0\right)}{m_1\left(y\right)}.$$

4. Análisis de regresión lineal

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

En el presente capítulo, interesa describir la distribución condicional de una variable aleatoria Y para valores dados de algunas otras variables $X_1, ..., X_k$. Los valores de las variables aleatorias $X_1, ..., X_k$ pueden ser observados en un experimento junto con los de Y o pueden ser variables de control cuyos valores van a ser determinados por el experimentador, además, se puede estudiar la distribución condicional de Y dadas $X_1, ..., X_k$.

Definición 4.1. *Se denomina* ***función de regresión*** *de Y sobre $X_1, ..., X_k$, o simplemente* ***regresión*** *de Y sobre $X_1, ..., X_k$ a la esperanza condicional de Y dados los valores de $X_1 = x_1, ..., X_k = x_k$, es decir, $E[Y|x_1, ..., x_k]$*

Supongamos que la función de regresión $E[Y|x_1, ..., x_k]$ es una función lineal con la siguiente forma

$$E[Y|x_1, ..., x_k] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Los coeficientes $\beta_0, \beta_1, ..., \beta_k$ se llaman **coeficientes de regresión**, los cuales son desconocidos y se consideran parámetros cuyos valores se estiman.

4.1. Regresión lineal simple

En el modelo de regresión lineal simple, se tiene un sólo regresor X que tiene una relación con una variable respuesta Y, donde la relación es una recta. Este modelo es de la forma,

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

donde la ordenada al origen β_0 y la pendiente β_1 son constantes desconocidas y ε es un componente aleatorio del error.

Supongamos que los errores $\varepsilon_i \sim N(0, \sigma^2)$ y no están correlacionados; en consecuencia las variables aleatorias $Y_1, ..., Y_n$ son independientes.

Considere que el regresor X se tiene como dato, y su error es despreciable, mientras que la respuesta Y es una variable aleatoria, se tiene entonces una distribución de probabilidad de Y para cada valor posible de X, así la media y la varianza de la distribución son

$$\begin{aligned} E[Y|X = x] &= E[\beta_0 + \beta_1 x + \varepsilon] \\ &= \beta_0 + \beta_1 x \\ Var[Y|X = x] &= \sigma^2 \end{aligned}$$

A un problema como el descrito anteriormente es llamado modelo de **regresión lineal simple**, el término de simple se emplea por que se está considerando la regresión de Y sobre una sola variable X.

4.1.1. Función de verosimilitud y estimación

Como los parámetros β_0 y β_1 son desconocidos, se deben estimar con los datos de la muestra. Supongamos que se tienen n pares de observaciones (x_i, y_i) con $i = 1, ..., n$, además sabemos que para cualesquiera $x_1, ..., x_n$ las variables aleatorias $Y_1, ..., Y_n$ son independientes y $Y_i \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$ para $i = 1, ..., n$.
La función de verosimilitud de β_0, β_1 y σ^2es:

$$L\left(y|x, \beta_0, \beta_1, \sigma^2\right) = \frac{1}{(2\pi)^{\frac{n}{2}}\,\sigma^n} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \quad (4.1)$$

Utilizamos ahora 4.1 para encontrar los estimadores máximo verosímiles que son aquellos que maximizan $L\left(y|x, \beta_0, \beta_1, \sigma^2\right)$. También sabemos en general que si x_0 es un punto crítico de $f(x)$ entonces x_0 es punto crítico de $\ln(f(x))$. Por lo que será equivalente hallar los máximos de $L\left(y|x, \beta_0, \beta_1, \sigma^2\right)$ que los de $\ln(L\left(y|x, \beta_0, \beta_1, \sigma^2\right))$. A esta última transformación de la verosimilitud la llamaremos log-verosimilitud y la denotaremos por $l\left(y|x, \beta_0, \beta_1, \sigma^2\right)$. Así procedemos a hallar la log-verosimilitud

$$\begin{aligned} l\left(y|x, \beta_0, \beta_1, \sigma^2\right) &= \ln\left(L\left(y|x, \beta_0, \beta_1, \sigma^2\right)\right) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

Calculamos la derivada parcial de $l\left(y|x, \beta_0, \beta_1, \sigma^2\right)$ con respecto a β_0

$$\frac{\partial}{\partial \beta_0}\left(l\left(y|x, \beta_0, \beta_1, \sigma^2\right)\right) = \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i$$

Ahora igualamos a cero

$$\begin{aligned} \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i &= 0 \\ \beta_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\beta_1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Por tanto

$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \tag{4.2}$$

Por otro lado calculamos la derivada parcial de $l\left(y|x, \beta_0, \beta_1, \sigma^2\right)$ con respecto a β_1

$$\frac{\partial}{\partial \beta_1}\left(l\left(y|x, \beta_0, \beta_1, \sigma^2\right)\right) = \sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2$$

Igualando a cero

$$\begin{aligned}
\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} y_i x_i - \beta_0 n\overline{x} - \beta_1 \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} y_i x_i - (\overline{y} - \beta_1 \overline{x}) n\overline{x} - \beta_1 \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} y_i x_i - n\overline{yx} + n\beta_1 \overline{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 &= 0 \\
\sum_{i=1}^{n} y_i x_i - n\overline{yx} + \beta_1 (n\overline{x}^2 - \sum_{i=1}^{n} x_i^2) &= 0
\end{aligned}$$

Despejando $\beta_1$obtenemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{4.3}$$

Para facilitar la notación consideremos las siguientes igualdades

$$\begin{aligned}
S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 &= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \\
S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 &= \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \\
S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \overline{x}
\end{aligned}$$

de esta manera los estimadores son

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{4.4}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{4.5}$$

Notesé que se han realizado tres supuestos importantes sobre la distribución conjunta de $Y_1, ..., Y_n$ para cualesquiera valores de $x_1, ..., x_n$.

1. Normalidad. Se ha supuesto que cada variable aleatoria ε_i tiene una distribución normal y en consecuencia Y_i tiene distribución normal.
2. Independencia. Se ha supuesto que las variables aleatorias $\varepsilon_1, ..., \varepsilon_n$ son independientes y consecuentemente $Y_1, ..., Y_n$ son independientes.
3. Homocedasticidad. De manera implícita se ha supuesto que las variables aleatorias $\varepsilon_1, ..., \varepsilon_n$ tienen la misma varianza a saber σ^2.

4.1.2. Propiedades de los estimadores

Es importante mencionar que los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ son funciones lineales de $Y_1, ..., Y_n$, y dado que estas variables aleatorias son independientes e idénticamente distribuidas (tienen una distribución normal) , también $\hat{\beta}_0$ y $\hat{\beta}_1$tendrán una distribución gaussiana, solo basta determinar sus correspondientes medias y varianzas. Calculemos el valor esperado de $\hat{\beta}_1$

$$
\begin{aligned}
E\left[\hat{\beta}_1\right] &= E\left[\frac{S_{xy}}{S_{xx}}\right] \\
&= \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,E\left[y_i\right]}{S_{xx}} \\
&= \frac{\sum_{i=1}^{n}(x_i-\overline{x})}{S_{xx}}\left(\beta_0+\beta_1 x_i\right) \\
&= \beta_0\frac{\sum_{i=1}^{n}(x_i-\overline{x})}{S_{xx}}+\beta_1\frac{\sum_{i=1}^{n}(x_i-\overline{x})x_i}{S_{xx}} \\
&= \beta_0\frac{n\overline{x}-n\overline{x}}{S_{xx}}+\beta_1\frac{\sum_{i=1}^{n}x_i^2-\overline{x}x_i}{S_{xx}} \\
&= \beta_1\frac{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}{S_{xx}} \\
&= \beta_1\frac{S_{xx}}{S_{xx}}
\end{aligned}
$$

Por lo tanto

$$
E\left[\hat{\beta}_1\right] = \beta_1 \tag{4.6}
$$

Esto implica que $\hat{\beta}_1$ es un estimador insesgado de β_1. Por otro lado la varianza de $\hat{\beta}_1$ será

$$
\begin{aligned}
Var\left(\hat{\beta}_1\right) &= Var\left(\frac{S_{xy}}{S_{xx}}\right) \\
&= Var\left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})\,y_i}{S_{xx}}\right) \\
&= \frac{S_{xx}}{\left(S_{xx}\right)^2}\sigma^2 \\
&= \frac{1}{S_{xx}}\sigma^2
\end{aligned}
$$

Por lo que

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \tag{4.7}$$

En consecuencia $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$. Ahora el valor esperado y la varianza de $\hat{\beta}_0$ son

$$\begin{aligned} E\left[\hat{\beta}_0\right] &= E\left[\overline{y} - \hat{\beta}_1\overline{x}\right] \\ &= E\left[\overline{y}\right] - \overline{x}E\left[\hat{\beta}_1\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] - \overline{x}\beta_1 \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[y_i\right] - \overline{x}\beta_1 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i\right) - \overline{x}\beta_1 \\ &= \beta_0 + \overline{x}\beta_1 - \overline{x}\beta_1 \end{aligned}$$

De lo anterior podemos observar que $\hat{\beta}_0$ es un estimador insesgado de β_0.

$$E\left[\hat{\beta}_0\right] = \beta_0 \tag{4.8}$$

Note que $Var\left(\overline{y}\right) = \frac{\sigma^2}{n}$ y $Cov\left(\overline{y}, \hat{\beta}_1\right) = 0$, así se tiene que

$$\begin{aligned} Var\left(\hat{\beta}_0\right) &= Var\left(\overline{y} - \hat{\beta}_1\overline{x}\right) \\ &= Var\left(\overline{y}\right) + \overline{x}^2 Var\left(\hat{\beta}_1\right) - 2\overline{x}Cov\left(\overline{y}, \hat{\beta}_1\right) \\ &= Var\left(\overline{y}\right) + \overline{x}^2 Var\left(\hat{\beta}_1\right) \end{aligned}$$

$$Var\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) \tag{4.9}$$

Y consecuentemente $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$

Teorema de Gauss-Markov para el modelo de regresión lineal simple

Este teorema es un resultado importante de los estimadores máximo verosímiles de β_0 y β_1, establece que para el modelo $Y = \xi_0\beta_0 + \beta_1 X + \varepsilon$, con las hipótesis

usuales , $E[\varepsilon] = 0$, $Var(\varepsilon) = \sigma^2$y errores no correlacionados, los estimadores máximo verosímiles son insesgados y tienen varianza mínima en comparación con todos los demás estimadores insesgados que sean combinaciones lineales de $Y_1, ..., Y_n$.

Teorema 4.1. *Sea $\tau = \xi_0\beta_0 + \xi_1\beta_1 + \xi_2$, donde ξ_0, ξ_1, ξ_2 son constantes dadas. Entonces , entre todos los estimadores insesgados de τ que son una función lineal de las observaciones $Y_1, ..., Y_n$, el estimador $\hat{\tau} = \xi_0\hat{\beta}_0 + \xi_1\hat{\beta}_1 + \xi_2$ tiene la menor varianza para todos los posibles valores de los parámetros $\beta_0, \beta_1 y\ \sigma^2$.*

Ahora nos proponemos obtener la distribución conjunta de $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\sigma^2}$.
Para los valores $x_1, ..., x_n$ con $n \geq 3$ supongamos que $Y_1, ..., Y_n$ son independientes y que $Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$.
Sea

$$S_x = \left[\sum_{i=1}^{n} (x_i - \overline{x})^2\right]^{1/2}$$

Consideremos los siguientes vectores n-dimensionales $a_1 = (a_{11}, ..., a_{1n})$ y $a_2 = (a_{21}, ..., a_{2n})$, cuyas entradas estarán definidas de la siguiente manera

$$\begin{aligned} a_{1j} &= \left(\frac{1}{n}\right)^{1/2} \\ a_{2j} &= \frac{1}{S_x} \end{aligned}$$

para $j = 1, ...n$.
Notemos que

$$\begin{aligned} \sum_{j=1}^{n} a_{1j}^2 &= \sum_{j=1}^{n} \frac{1}{n} = 1 \\ \sum_{j=1}^{n} a_{2j}^2 &= \frac{(x_1 - \bar{x})^2}{S_{xx}} + ... + \frac{(x_n - \bar{x})^2}{S_{xx}} \\ &= \frac{S_{xx}}{S_{xx}} = 1 \\ \sum_{j=1}^{n} a_{1j}a_{2j} &= \left(\frac{1}{n}\right)^{\frac{1}{2}} \left(\frac{x_1 - \bar{x}}{S_x} + ... + \frac{x_n - \bar{x}}{S_x}\right) = 0 \end{aligned}$$

Ya que las entradas de los vectores a_1 y a_2 cumplen las propiedades antes mencionadas, se puede construir una matriz ortogonal $A \in M_{n\times n}(\mathbb{R})$, con a_1 y a_2 la primera y segunda fila de tal matriz respectivamente.

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \cdots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

Sea Z el vector aleatorio definido por medio de $Z = AY$, donde

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$$

La distribución conjunta de $Z_1, .., Z_n$ se puede encontrar con el siguiente teorema.

Teorema 4.2. *Supongamos que $Y_1, ..., Y_n$ son variables aleatorias independientes y que cada una de ellas tiene una distribución normal , tal que $Var(Y_i) = \sigma^2$. Si A es una matriz ortogonal de tamaño $n \times n$ y $Z = AY$, entonces las variables aleatorias $Z_1, ..., Z_n$ también son independientes y cada una tiene una distribución normal con $Var(Z_i) = \sigma^2$*

En un problema de regresión lineal simple, las observaciones $Y_1, ..., Y_n$ cumplen las condiciones del teorema anterior, por lo que las componentes del vector aleatorio $Z = AY$ serán independientes y cada una tendrá una distribución normal con varianza σ^2.

Las dos primeras componentes Z_1 y Z_2 de Z son:

$$Z_1 = \sum_{j=1}^{n} a_{1j} Y_j = \frac{1}{n^{1/2}} \sum_{j=1}^{n} Y_j = n^{1/2} \bar{Y} \tag{4.10}$$

Pero sabemos que $\hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta}_1$, así

$$Z_1 = n^{1/2} \left(\hat{\beta}_0 + \bar{X}\hat{\beta}_1 \right) \tag{4.11}$$

Por otra parte, como $\hat{\beta}_1 = \dfrac{\sum_{j=1}^{n} (x_j - \bar{x}) Y_j}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$, la segunda componente de la matriz A es :

$$
\begin{aligned}
Z_2 = \sum_{j=1}^{n} a_{2j} Y_j &= \frac{1}{S_x} \sum_{j=1}^{n} (x_j - \bar{x})\, Y_j \\
&= \frac{1}{S_x} \sum_{j=1}^{n} (x_j - \bar{x})^2\, \hat{\beta}_1 \\
&= \frac{S_x^2}{S_x} \hat{\beta}_1
\end{aligned}
$$

$$
Z_2 = S_x \hat{\beta}_1 \tag{4.12}
$$

Consideremos la variable aleatoria S^2 dada por :

$$
S^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \tag{4.13}
$$

Basta con mostrar que S^2 es independiente del vector aleatorio $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ para afirmar que $\hat{\sigma^2}$ es independiente de $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Podemos reescribir a S^2 usando la siguiente igualdad ,

$$
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}
$$

por tanto

$$
\begin{aligned}
S^2 &= \sum_{i=1}^{n} \left[\left(Y_i - \bar{Y} \right) - \hat{\beta}_1 \left(x_i - \bar{x} \right) \right]^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right) \left(x_i - \bar{x} \right) + \left(\hat{\beta}_1 \right)^2 S_x^2 \\
&= \sum_{i=1}^{n} Y_i^2 - 2\bar{Y} \sum_{i=1}^{n} Y_i + n\bar{Y}^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} \left(Y_i \right) \left(x_i - \bar{x} \right) - 2\hat{\beta}_1 \bar{Y} \sum_{i=1}^{n} \left(x_i - \bar{x} \right) + \left(\hat{\beta}_1 \right)^2 S_x^2
\end{aligned}
$$

adicionalmente sabemos que $\hat{\beta}_1 S_x^2 = \sum_{i=1}^{n} \left(x_i - \bar{x} \right) Y_i$, entonces

$$
\begin{aligned}
S^2 &= \sum_{i=1}^{n} Y_i^2 - 2\bar{Y} \sum_{i=1}^{n} Y_i + n\bar{Y}^2 - 2 \left(\hat{\beta}_1 \right)^2 S_x^2 + \left(\hat{\beta}_1 \right)^2 S_x^2 \\
&= \sum_{i=1}^{n} Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2 - 2 \left(\hat{\beta}_1 \right)^2 S_x^2 + \left(\hat{\beta}_1 \right)^2 S_x^2
\end{aligned}
$$

$$S^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 - \left(\hat{\beta}_1\right)^2 S_x^2 \tag{4.14}$$

Puesto que $Z = AY$, donde A es una matriz ortogonal, se cumple que

$$\begin{aligned}
\sum_{i=1}^{n} Y_i^2 &= Y^t Y \\
&= Y^t I Y \\
&= Y^t A^t A Y \\
&= Z^t Z \\
&= \sum_{i=1}^{n} Z_i^2
\end{aligned}$$

De las ecuaciones 4.10 y 4.12, obtenemos que

$$\begin{aligned}
S^2 &= \sum_{i=1}^{n} Z_i^2 - Z_1^2 - Z_2^2 \\
&= \sum_{i=3}^{n} Z_i^2
\end{aligned}$$

Note que $\{Z_i\}_{i=1}^n$ es una sucesión de variables aleatorias independientes y como S^2 es únicamente la suma de los cuadrados de las variables $Z_3, ..., Z_n$, entonces S^2 y el vector (Z_1, Z_2) son independientes, pero $\hat{\beta}_0$ y $\hat{\beta}_1$ son funciones unicamente de Z_1 y Z_2,

$$\begin{aligned}
\hat{\beta}_0 &= \frac{Z_1}{n^{1/2}} - \frac{\bar{x} Z_2}{S_x} \\
\hat{\beta}_1 &= \frac{Z_2}{S_x}
\end{aligned}$$

por lo tanto podemos concluir que S^2 y $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ son independientes.

Por otro lado tenemos que para $i = 3, ..., n$

$$Z_i = \sum_{j=1}^{n} a_{ij} Y_j$$

de esta forma

$$\begin{aligned} E(Z_i) &= \sum_{j=1}^{n} a_{ij} E(Y_j) \\ &= \sum_{j=1}^{n} a_{ij} (\beta_0 + \beta_1 x_j) \\ &= \sum_{j=1}^{n} a_{ij} [\beta_0 + \beta_1 \bar{x} + \beta_1 (x_j - \bar{x})] \end{aligned}$$

$$E(Z_i) = (\beta_0 + \beta_1 \bar{x}) \sum_{j=1}^{n} a_{ij} + \beta_1 \sum_{j=1}^{n} a_{ij} (x_j - \bar{x}) \tag{4.15}$$

Ya que A es una matriz ortogonal, la suma de los productos de los correspondientes términos en dos filas cualesquiera debe ser cero. En particular para $i = 3, ..., n$

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} a_{1j} &= 0 \\ \sum_{j=1}^{n} a_{ij} a_{2j} &= 0 \end{aligned}$$

Por la forma de a_{1j} y a_{2j} se tiene que

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} &= 0 \\ \sum_{j=1}^{n} a_{ij} (x_j - \bar{x}) &= 0 \end{aligned}$$

Por lo tanto $E(Z_i) = 0$ para $i = 1, ..., n$

De acuerdo a todo lo anterior resulta que las $n-2$ variables aleatorias $Z_3, ..., Z_n$ son independientes y tienen una distribución normal con media 0 y varianza σ^2, de acuerdo al teorema anterior.

Ya que $S^2 = \sum_{i=3}^{n} Z_i^2$, la variable aleatoria $\frac{S^2}{\sigma^2}$ tendrá una distribución χ^2 con $n-2$ grados de libertad. Consideremos el estimador máximo verosímil de σ^2, $\hat{\sigma^2} = \frac{S^2}{n}$ por lo que $S^2 = n\hat{\sigma^2}$ y de los resultados anteriores se deduce que $\hat{\sigma^2}$ es independiente de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$y finalmente

$$\frac{n\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2} \tag{4.16}$$

4.1.3. Contraste de hipótesis sobre β_0

Se desea contrastar

$$\begin{aligned} H_0 &: \beta_0 = \beta_0^* \\ H_1 &: \beta_0 \neq \beta_0^* \end{aligned}$$

con $\beta_0^* \in \mathbb{R}$. De las ecuaciones 4.8 y 4.9 cuando la hipótesis H_0 es cierta , la siguiente variable aleatoria tendrá una distribución normal estándar.

$$U = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}}$$

Se desea construir un contraste para que H_0 se rechace cuando β_0 se encuentre lejos del valor de β_0^* y se acepte en otro caso. Al ser σ desconocido no es posible basarse solo en el estadístico U. Como $\hat{\beta}_0$ y S^2 son variables aleatorias independientes tenemos que U y S^2 también serán independientes. De esta manera bajo H_0 se tiene que

$$\begin{aligned} T &= \frac{\frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}}}{\sqrt{\left(\frac{n\hat{\sigma^2}}{\sigma^2}\right)/(n-2)}} \\ &= \sqrt{\frac{n-2}{n}} \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma^2}}} \end{aligned}$$

Por tanto

$$T \sim t_{n-2}$$

De acuerdo al estadístico anterior, se rechaza H_0 si $|T| > c$ donde c es una constante cuyo valor se elige para cualquier $\alpha \in (0, 1)$. De manera formal c es algún cuantil de la distribución encontrada anteriormente.

4.1.4. Contraste de hipótesis sobre β_1

Para hacer un contraste con respecto a β_1, con $\beta_1^* \in \mathbb{R}$

$$\begin{aligned} H_0 &: \beta_1 = \beta_1^* \\ H_1 &: \beta_1 \neq \beta_1^* \end{aligned}$$

Bajo H_0 la variable aleatoria

$$V = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

por tanto el estadístico que utilizaremos para este contraste es

$$\begin{aligned} U &= \frac{\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}}}{\sqrt{\left(\frac{n\hat{\sigma^2}}{\sigma^2}\right)/(n-2)}} \\ &= \sqrt{\frac{n-2}{n}} \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S_{xx}\hat{\sigma^2}}} \end{aligned}$$

Por tanto

$$U \sim t_{n-2}$$

Entonces la hipótesis H_0 se rechaza si $|U| > c$ con c una constante apropiada cuyo valor se puede elegir para algún valor de significancia $0 < \alpha < 1$. Análogamente c es algún cuantil de la distribución encontrada anteriormente.

4.2. Regresión lineal múltiple

Supóngase que se tienen n vectores de observaciones. Para $i = 1, ..., n$ sea $(x_{i1}, x_{i2}, ..., x_{ik})$ el vector de valores observados , nuestro objetivo será encontrar una función que se ajuste de la "mejor" manera a estas observaciones y una alternativa es la propuesta del modelo de regresión lineal.

Un modelo de regresión lineal en donde intervienen más de una variable regresora se llama modelo de regresión múltiple, si el modelo es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

entonces se conoce como modelo de regresión múltiple, con k regresores, los parámetros β_j , $j = 0, 1, ..., k$ se conocen como coeficientes de regresión.

Este modelo describe un hiperplano en el espacio vectorial $\mathbb{R}^k$. Entonces si, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ y además los errores no están correlacionados y si se supone que la función de regresión es una función lineal, entonces

$$E[Y|X_1, ..., X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

El parámetro β_j representa el cambio esperado en la respuesta Y por cambio en la variable regresora X_j , cuando fijamos $X_i \ \forall i \neq j$

4.2.1. Estimación de los coeficientes de regresión

Supóngase que se tienen $n > k$ observaciones y sea y_i la $i - ésima$ respuesta observada y sea x_{ij} el $i - ésimo$ valor de la variable regresora X_j.

$$\begin{array}{ccccc} i & y & x_1 & \cdots & x_k \\ 1 & y_1 & x_{11} & \cdots & x_{1k} \\ 2 & y_2 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ n & y_n & x_{n1} & \cdots & x_{nk} \end{array}$$

Entonces la $i - esima$ observación satisface

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \ i = 1, ..., n$$

y la suma de los cuadrados de los errores

$$S = \sum_{j=1}^{k} \varepsilon_i^2 = \sum_{j=1}^{k} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

La forma más sencilla de realizar la estimación por mínimos cuadrados es escribiéndolo en forma matricial, de la siguiente forma:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}_{n\times r},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix}_{r\times 1}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1},$$

donde $r = k + 1$ es el número de parámetros a determinar.

De esta forma el modelo y su correspondiente suma de los cuadrados de los errores pueden ser expresados en la siguiente forma

$$Y = X\beta + \varepsilon$$

Definimos en este caso a S como $S = \varepsilon^t\varepsilon$

$$\begin{aligned} S &= (Y - X\beta)^t (Y - X\beta) \\ &= Y^tY - 2\beta^tX^tY + \beta^tX^tX\beta \end{aligned}$$

Hemos puesto especial importancia en la suma de los cuadrados de los errores pues nos será de utilidad para poder estimar los coeficientes de regresión.

Calculamos $\frac{\partial S}{\partial \beta}$

$$\frac{\partial S}{\partial \beta} = -2X^tY + 2X^tX\beta$$

Si X^tX es no singular entonces

$$\hat{\beta} = (X^tX)^{-1}X^tY$$

En el caso en el que X^tX fuera singular (esto sucede cuando se tiene multicolinealidad), para resolver el problema se tendría que calcular la matriz pseudoinversa, lo cual no será desarrollado en esta tesis. En adelante se supondrá que X^tX es no singular.

4.2.2. Estimación de σ^2

En el modelo $Y = X\beta + \varepsilon$ los errores se distribuyen normal , además de ser independientes , con media cero y varianza constante e igual a σ^2, es decir, $\varepsilon \sim N_n(\overline{0}, \sigma^2 I)$ donde I es la matriz identidad en el espacio $M_{n\times n}(\mathbb{R})$, de esta forma

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\sigma^2}} exp\left\{-\frac{\varepsilon_i^2}{2\sigma^2}\right\}$$

Y la función conjunta y la función de verosimilitud son

$$f(\varepsilon_1, ..., \varepsilon_n) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp\left\{-\frac{\varepsilon^t\varepsilon}{2\sigma^2}\right\}$$

$$L(Y, X, \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp\left\{-\frac{(Y - X\beta)^t (Y - X\beta)}{2\sigma^2}\right\}$$

De esta forma la log-verosimilitud esta dada por

$$l\left(Y, X, \beta, \sigma^2\right) = -\frac{n}{2}\ln(2\pi) - n\ln\left(\sigma^2\right) - \frac{(Y - X\beta)^t (Y - X\beta)}{2\sigma^2}$$

para un valor fijo de σ , $l\left(Y, X, \beta, \sigma^2\right)$ se maximiza al minimizar $(Y - X\beta)^t (Y - X\beta)$

$$\frac{\partial l\left(Y, X, \beta, \sigma^2\right)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{(Y - X\beta)^t (Y - X\beta)}{\sigma^3} = 0$$

Por tanto el estimador máximo verosímil para σ^2 es

$$\hat{\sigma^2} = \frac{(Y - X\beta)^t (Y - X\beta)}{n}$$

Es importante mencionar que el estimador por mínimos cuadrados ordinarios para β es igual al estimador máximo verosímil.

Estimador insesgado de σ^2

Ahora encontraremos un estimador insesgado para σ^2, usando la suma de cuadrados de los residuales(SCR). Definida como:

$SCR = e^t e = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$, donde $e = Y - X\hat{\beta}$

Desarrollemos el siguiente producto

$$\begin{aligned} e^t e &= \left(Y - X\hat{\beta}\right)^t \left(Y - X\hat{\beta}\right) \\ &= Y^t Y - Y^t X\hat{\beta} - \hat{\beta}^t X^t X\hat{\beta} + \hat{\beta}^t X^t X\hat{\beta} \\ &= Y^t Y - 2\hat{\beta}^t X^t Y + \hat{\beta}^t X^t X\hat{\beta} \end{aligned}$$

Como $X^t X\hat{\beta} = X^t Y$

$$SCR = Y^t Y - \hat{\beta}^t X^t Y$$

Cabe mencionar que la suma de cuadrados de los residuales tiene $n - r$ grados de libertad asociados a ella, ya que se estiman r parámetros en el modelo de regresión. Para percatarnos de ello es conveniente reescribir la suma de cuadrados de los residuales de la siguiente manera:

$$\begin{aligned} SCR &= \left[Y - X(X^t X)^{-1} X^t Y\right]^t \left[Y - X(X^t X)^{-1} X^t Y\right] \\ &= Y^t \left[I - X(X^t X)^{-1} X^t\right] Y \end{aligned}$$

con I la matriz identidad en el espacio $M_{n\times n}(\mathbb{R})$, además note que la matriz $I - X(X^t X)^{-1} X^t$ es idempotente y simétrica pues

$$\begin{aligned} I - X(X^t X)^{-1} X^t &= \left[I - X(X^t X)^{-1} X^t\right]^2 \text{y} \\ I - X(X^t X)^{-1} X^t &= \left[I - X(X^t X)^{-1} X^t\right]^t \end{aligned}$$

por lo que

$$I - X(X^t X)^{-1} X^t = \left[I - X(X^t X)^{-1} X^t\right]^t \left[I - X(X^t X)^{-1} X^t\right]$$

Para llegar a la siguiente igualdad resaltaremos que $I \in M_{n\times n}(\mathbb{R})$ y $X \in M_{n\times r}(\mathbb{R})$

$$\begin{aligned} tr\left[I - X(X^t X)^{-1} X^t\right] &= tr\left[I\right] - tr\left[X(X^t X)^{-1} X^t\right] \\ &= tr[I_{n\times n}] - tr\left[I_{r\times r}\right] = n - r \end{aligned}$$

Ya que $I - X(X^t X)^{-1} X^t$ es idempotente se tiene que

$$rango\left[I - X(X^t X)^{-1} X^t\right] = tr\left[I - X(X^t X)^{-1} X^t\right] = n - r$$

Por el Teorema $B.3$ del Apéndice

$$\frac{SCR}{\sigma^2} = \frac{1}{\sigma^2} Y^t \left[I - X(X^t X)^{-1} X^t \right] Y \sim \chi^2_{n-r,\lambda}$$

El parámetro de no centralidad λ es:

$$\lambda = \frac{1}{\sigma^2} E[Y]^t \left[I - X(X^t X)^{-1} X^t \right] E[Y]$$

Pero $E[Y] = X\beta$, entonces

$$\begin{aligned} \lambda &= \frac{1}{\sigma^2} \beta^t X^t \left[I - X(X^t X)^{-1} X^t \right] X\beta \\ &= \frac{1}{\sigma^2} \beta^t \left[X^t X - X^t X \left(X^t X \right)^{-1} X^t X \right] \beta \end{aligned}$$

Por lo que

$$\lambda = 0$$

Y por lo tanto

$$\tfrac{SCR}{\sigma^2} \sim \chi^2_{n-r}$$

4.2.3. Propiedades de los estimadores

Supongamos que Y es un vector aleatorio de tamaño n, a saber $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$

Definición 4.2. *La esperanza $E[Y]$del vector aleatorio Y, se define como el vector n-dimensional cuyas componentes son las esperanzas de las componentes individuales de Y. Así*

$$E[Y] = \begin{bmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{bmatrix}$$

Definición 4.3. *La matriz de covarianzas del vector aleatorio Y se define como la matriz cuadrada de $n \times n$ tal que la entrada (i, j) sea $Cov(Y_i, Y_j)$ $\forall i, j$. Si $Cov(Y_i, Y_j) = \sigma_{ij}$ entonces*

$$Cov\left(Y\right)=\left(\begin{array}{ccc}\sigma_{11} & \cdots & \sigma_{1n}\\ \vdots & \ldots & \vdots\\ \sigma_{n1} & \cdots & \sigma_{nn}\end{array}\right)$$

Notemos que $Var\left(Y_i\right)=Cov\left(Y_i,Y_i\right)=\sigma_{ii}$ y por tanto en la diagonal de la matriz de covarianza tendremos las varianzas de $Y_1,...,Y_n$. Además como $Cov\left(Y_i,Y_j\right)=Cov\left(Y_j,Y_i\right)$ la matriz $Cov\left(Y\right)$ es simétrica.

De acuerdo a las definiciones anteriores , como $E\left(\varepsilon\right)=0$, $Var\left(\varepsilon\right)=\sigma^2$ y $Y=X\beta+\varepsilon$

$$E\left[Y\right]=E\left[X\beta+\varepsilon\right]=X\beta+E\left[\varepsilon\right]$$

$$E\left[Y\right]=X\beta$$

Sabemos que las variables aleatorias $Y_1,...,Y_n$ son independientes y la varianza de cada una es σ^2, por tanto

$$Cov\left(Y\right)=\sigma^2I$$

Para encontrar las esperanzas y covarianzas de los estimadores, enunciamos el siguiente teorema.

Teorema 4.3. *Supongamos que Y es un vector aleatorio n-dimensional, para el cual existe $E\left[Y\right]$ y $Cov\left(Y\right)$. Supongamos también que A es una matriz de tamaño $p\times n$ cuyos elementos son constantes y que v es un vector aleatorio p-dimensional definido como $v=AY$.Entonces*

$$E\left[v\right]=AE\left[Y\right]\ y\ Cov\left(v\right)=ACov\left(Y\right)A^t$$

Demostración

Sea

$$A \quad = \quad \left[\begin{array}{ccc}a_{11} & \cdots & a_{1n}\\ \vdots & \ldots & \vdots\\ a_{p1} & \cdots & a_{pn}\end{array}\right]$$

la i-ésima componente de v es

$$v_i \quad = \quad \sum_{j=1}^{n}a_{ij}Y_j$$

y la i-ésima componente de $E\left(v\right)$ es

$$E[v_i] = E\left[\sum_{j=1}^{n} a_{ij}Y_j\right] = \sum_{j=1}^{n} a_{ij}E[Y_j]$$

sin embargo esta es la i-ésima componente del vector $AE[Y]$, por tanto

$$E[v] = AE[Y].$$

La matriz de covarianzas de V es una matriz de tamaño $p \times p$. Para $h, k = 1, ...p$

$$\begin{aligned} Cov(V_h, V_k) &= Cov\left(\sum_{r=1}^{n} a_{hr}Y_r, \sum_{s=1}^{n} a_{ks}Y_s\right) \\ &= \sum_{r=1}^{n}\sum_{s=1}^{n} a_{hr}a_{ks}Cov(Y_r, Y_s) \end{aligned}$$

Pero este elemento es la entrada (r, s) de la matriz $ACov(Y)A^t$, entonces

$$Cov(V) = ACov(Y)A^t$$

■

Utilizando el teorema anterior, el valor esperado de $\hat{\beta}$ es

$$\begin{aligned} E\left[\hat{\beta}\right] &= E\left[\left(X^tX\right)^{-1}X^tY\right] \\ &= E\left[\left(X^tX\right)^{-1}X^t(X\beta + \varepsilon)\right] \\ &= E\left[\left(X^tX\right)^{-1}X^tX\beta + \left(X^tX\right)^{-1}X^t\varepsilon\right] \\ &= E[\beta] + \left(X^tX\right)^{-1}X^tE[\varepsilon] \\ &= \beta \end{aligned}$$

$$E\left[\hat{\beta}\right] = \beta \tag{4.17}$$

Por tanto el estimador es insesgado y

$$E\left[\hat{\beta}_j\right] = \beta_j \ \forall j$$

Utilizando nuevamente el teorema anterior y recordando que

$$\hat{\beta} = \left(X^t X\right)^{-1} X^t Y$$

y note que $\left(X^t X\right)^{-1} X^t$ es una matriz de constantes, por tanto

$$\begin{aligned} Cov\left(\hat{\beta}\right) &= \left[\left(X^t X\right)^{-1} X^t\right] Cov\left(Y\right) \left[\left(X^t X\right)^{-1} X^t\right]^t \\ &= \left(X^t X\right)^{-1} X^t \left(\sigma^2 I\right) X \left(X^t X\right)^{-1} \\ &= \left(\sigma^2 I\right) \left(X^t X\right)^{-1} X^t X \left(X^t X\right)^{-1} \\ &= \sigma^2 \left(X^t X\right)^{-1} \end{aligned}$$

$$Cov\left(\hat{\beta}\right) = \sigma^2 \left(X^t X\right)^{-1} \tag{4.18}$$

Sea $A = \left(X^t X\right)^{-1}$, entonces para $j = 1, ..., n$

$$Var\left(\hat{\beta}_j\right) = \sigma^2 a_{jj}$$

Para $i \neq j$

$$Cov\left(\hat{\beta}_i, \hat{\beta}_j\right) = \sigma^2 a_{ij}$$

Teorema de Gauss-Markov para el caso general

Es importante poder garantizar que el estimador de β tenga varianza mínima y a través del teorema de Gauss-Markov lo podemos asegurar. Tengamos en cuenta los siguientes supuestos

- $E\left[Y_i\right] = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$
- $Y_1, ..., Y_n$ son no correlacionadas
- $Var\left(Y_i\right) = \sigma^2$ para $i = 1, ..., n$

Teorema 4.4. ***De Gauss-Markov***
Si $E\left(Y\right) = X\beta$ y $Cov\left(Y\right) = \sigma^2 I$, entre todos los estimadores insesgados de β que son combinaciones lineales de las observaciones $Y_1, ..., Y_n$, el estimador $\hat{\beta}$ tiene la menor varianza para todos los posibles valores de $\beta_0, ..., \beta_k$ y σ^2.

4.2.4. Pruebas de hipótesis

Supongamos que deseamos contrastar la hipótesis que uno de los coeficientes de la regresión β_j tiene un valor particular β_j^* , es decir:

$$\begin{aligned} H_0 &: \beta_j = \beta_j^* \\ H_1 &: \beta_j \neq \beta_j^* \end{aligned}$$

Como $Var\left(\hat{\beta}_j\right) = a_{jj}\sigma^2$, cuando H_0 es verdadera , la variable aleatoria W_j tiene una distribución normal estándar, donde $W_j$esta definida de la siguiente manera

$$W_j = \frac{\hat{\beta}_j - \beta_j^*}{a_{jj}^{1/2}\sigma}$$

Además recordemos que la variable aleatoria $\frac{SCR}{\sigma^2}$ tiene una distribución χ^2 con $n-r$ grados de libertad , al ser SCR independiente de $\hat{\beta}_j$, cuando H_0 es cierta tenemos que:

$$T_j = \frac{W_j}{\sqrt{\frac{1}{n-r}\left(\frac{SCR}{\sigma}\right)^2}} \quad \sim \quad t_{n-r}$$

4.3. Análisis Bayesiano del modelo de regresión lineal

4.3.1. Distribución a posteriori con distribución a priori no informativa

El modelo Bayesiano de regresión lineal múltiple es similar al modelo clásico de regresión lineal múltiple, excepto que en el primero se incluyen especificaciones sobre la distribución a priori de los parámetros. Del modelo de regresión múltiple

$$Y = X\beta + \varepsilon$$

en donde $\varepsilon_1, ..., \varepsilon_n$ son independientes e idénticamente distribuidos ; $\varepsilon_i \sim N\left(0, \sigma^2\right)$ entonces $\varepsilon \sim N_n(\overline{0}, \sigma^2 I)$ se obtuvo la función de verosimilitud

$$L(\beta, \sigma^2|Y) \quad = \quad \frac{1}{(2\pi)^{n/2}\,\sigma^n} exp\left[-\frac{1}{2\sigma^2}\left(Y - X\beta\right)^t\left(Y - X\beta\right)\right] \qquad (4.19)$$

Pero dado que $\hat{\beta} = (X^t X)^{-1} X^t Y$ y $\hat{Y} = X\hat{\beta}$

$$\begin{aligned}
(Y - X\beta)^t (Y - X\beta) &= (Y - X\beta)^t (Y - X\beta) - 2Y^t\hat{Y} + 2Y^t\hat{Y} \\
&= Y^tY - Y^tX\beta - \beta^tX^tY + \beta^tX^tX\beta - 2Y^tX\hat{\beta} + 2Y^tX\hat{\beta} \\
&= Y^tY - 2\beta X^tY + \beta^tX^tX\beta - 2Y^tX\hat{\beta} + 2Y^tX\hat{\beta} \\
&= Y^tY + \beta^tX^tX\beta - 2\beta^tX^tX(X^tX)^{-1}X^tY - 2Y^tX\hat{\beta} + \\
&\quad +2Y^tX(X^tX)^{-1}X^tX\hat{\beta} \\
&= Y^tY + \beta^tX^tX\beta - 2\beta^tX^tX\hat{\beta} - 2Y^tX\hat{\beta} + 2\hat{\beta}^tX^tX\hat{\beta} \\
&= Y^tY + \beta^tX^tX\beta - \beta^tX^tX\hat{\beta} - \hat{\beta}X^tX\beta - Y^tX\hat{\beta} - \hat{\beta}^tX^tY + \\
&\quad \hat{\beta}^tX^tX\hat{\beta} + \hat{\beta}^tX^tX\hat{\beta} \\
&= Y^tY - Y^tX\hat{\beta} - \hat{\beta}^tX^tY + \hat{\beta}^tX^tX\hat{\beta} + \beta^tX^tX\beta - \beta^tX^tX\hat{\beta} - \\
&\quad \hat{\beta}^tX^tX\beta + \hat{\beta}^tX^tX\hat{\beta} \\
&= \left(Y - X\hat{\beta}\right)^t \left(Y - X\hat{\beta}\right) + \left(\beta - \hat{\beta}\right)^t X^tX \left(\beta - \hat{\beta}\right) \\
&= (n-r)\left(\frac{1}{n-r}\right)\left(Y - \hat{Y}\right)^t \left(Y - \hat{Y}\right) + \left(\beta - \hat{\beta}\right)^t X^tX \left(\beta - \hat{\beta}\right)
\end{aligned}$$

De la igualdad anterior se puede reescribir a $(Y - X\beta)^t (Y - X\beta)$ como:

$$(Y - X\beta)^t (Y - X\beta) = vS^2 + \left(\beta - \hat{\beta}\right)^t X^tX \left(\beta - \hat{\beta}\right) \qquad (4.20)$$

con

$$S^2 = \frac{\left(Y - \hat{Y}\right)^t \left(Y - \hat{Y}\right)}{v} \text{ , } v = n - r$$

Al sustituir 4.20 en 4.19 obtenemos una nueva expresión para la verosimilitud

$$\begin{aligned}
L(\beta, \sigma^2|Y) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \times \\
&= \times exp\left\{-\frac{1}{2\sigma^2}\left[vS^2 + \left(\beta - \hat{\beta}\right)^t X^tX \left(\beta - \hat{\beta}\right)\right]\right\} \quad (4.21)
\end{aligned}$$

Como β es un parámetro de posición y σ^2 un parámetro de escala, las distribuciones a priori no informativas son:

$$\pi(\beta) \propto 1 \text{ y } \pi\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$$

Suponiendo que β y σ^2 son independientes se tiene que

$$\pi\left(\beta, \sigma^2\right) = \pi(\beta)\,\pi\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$$

Por el teorema de Bayes, la distribución a posteriori $\pi\left(\beta,\sigma^2|Y\right) \propto \pi\left(\beta,\sigma^2\right) L\left(\beta,\sigma^2|Y\right)$
En consecuencia

$$\begin{aligned}\pi\left(\beta,\sigma^2|Y\right) &= \frac{1}{\sigma^{n+2}} \times \\ &\times \quad exp\left\{-\frac{1}{2\sigma^2}\left[vS^2+\left(\beta-\hat{\beta}\right)^t X^t X\left(\beta-\hat{\beta}\right)\right]\right\} \end{aligned} \tag{4.22}$$

La función anterior puede ser representada por el siguiente producto

$$\pi\left(\beta,\sigma^2|Y\right) = \pi\left(\sigma^2|Y\right)\pi\left(\beta|Y,\sigma^2\right) \tag{4.23}$$

Distribución a posteriori para σ^2

$$\begin{aligned}\pi\left(\sigma^2|Y\right) &= \int_{-\infty}^{\infty}\pi\left(\beta,\sigma^2|Y\right)d\beta \\ &= \int_{-\infty}^{\infty}\frac{1}{\sigma^{n+2}}exp\left\{-\frac{1}{2\sigma^2}\left[vS^2+\left(\beta-\hat{\beta}\right)^t X^t X\left(\beta-\hat{\beta}\right)\right]\right\}d\beta \\ &= \frac{1}{\sigma^{n+2}}exp\left\{-\frac{vS^2}{2\sigma^2}\right\}\int_{-\infty}^{\infty}exp\left\{-\frac{1}{2}\left(\beta-\hat{\beta}\right)^t\frac{X^tX}{\sigma^2}\left(\beta-\hat{\beta}\right)\right\}d\beta \\ &= \frac{1}{\sigma^{n+2}}exp\left\{-\frac{vS^2}{2\sigma^2}\right\}\left(\sqrt{2\pi}\right)^r det\left(\frac{\left(X^tX\right)^{-1}}{\sigma^2}\right)^{1/2} \\ &\propto \frac{1}{\sigma^{n+2}}exp\left\{-\frac{vS^2}{2\sigma^2}\right\}\left(\sqrt{2\pi}\right)^r\frac{1}{\left[\left(\sigma^2\right)^{2r}\right]^{-1/2}} \\ &\propto \frac{1}{\sigma^{n+2-2r}}exp\left\{-\frac{vS^2}{2\sigma^2}\right\} \\ &= \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}+1-\frac{r}{2}}}exp\left\{-\frac{vS^2}{2\sigma^2}\right\} \\ &= \frac{1}{\left(\sigma^2\right)^{\frac{n-r}{2}+1}}exp\left\{-\frac{vS^2}{2\sigma^2}\right\} \\ &= \frac{1}{\left(\sigma^2\right)^{\frac{v}{2}+1}}exp\left\{-\frac{vS^2}{2\sigma^2}\right\}\end{aligned}$$

De lo anterior podemos observar que la distribución marginal posterior para σ^2 es el kernel de una distribución gamma inversa. Por lo tanto

$$\pi\left(\sigma^2|Y\right) = \left(\frac{vS^2}{2}\right)^{\frac{v}{2}}\frac{1}{\Gamma\left(\frac{v}{2}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{v}{2}+1}exp\left\{-\frac{vS^2}{2\sigma^2}\right\} \tag{4.24}$$

es decir,

$$\pi\left(\sigma^2|Y\right) \quad \sim \quad IGa\left(\frac{v}{2}, \frac{vS^2}{2}\right) \tag{4.25}$$

La distribución condicional a posteriori

$$\begin{aligned}
\pi\left(\beta|\sigma^2, Y\right) &= \frac{\pi\left(\beta, \sigma^2|Y\right)}{\pi\left(\sigma^2|Y\right)} \\
&\propto \frac{\frac{1}{\sigma^{n+2}} exp\left\{-\frac{1}{2\sigma^2}\left[vS^2 + \left(\beta-\hat{\beta}\right)^t X^t X\left(\beta-\hat{\beta}\right)\right]\right\}}{\left(\frac{1}{\sigma^2}\right)^{\frac{v}{2}+1} exp\left\{-\frac{vS^2}{2\sigma^2}\right\}} \\
&= \frac{1}{\sigma^r} exp\left\{-\frac{1}{2\sigma^2}\left(\beta-\hat{\beta}\right)^t X^t X\left(\beta-\hat{\beta}\right)\right\}
\end{aligned}$$

donde r es el número de parámetros.
De esta manera

$$\pi\left(\beta|\sigma^2, Y\right) \quad \sim \quad N_r\left(\hat{\beta}, \left(X^t X\right)^{-1}\sigma^2\right) \tag{4.26}$$

Si integramos con respecto a σ^2 la distribución a posteriori de la ecuación 4.22 obtendremos la distribución a posteriori marginal de β dada por:

$$\begin{aligned}
\pi\left(\beta|Y\right) &= \int_0^\infty \left(\frac{1}{\sigma^{n+2}}\right) \times \\
&\left(exp\left\{-\frac{1}{2\sigma^2}\left[vS^2 + \left(\beta-\hat{\beta}\right)^t X^t X\left(\beta-\hat{\beta}\right)\right]\right\}\right) d\sigma^2
\end{aligned} \tag{4.27}$$

Sea $z = \frac{vS^2 + \left(\beta-\hat{\beta}\right)^t X^t X\left(\beta-\hat{\beta}\right)}{2}$, de esta forma podemos reescribir 4.27 como:

$$
\begin{aligned}
\pi\left(\beta|Y\right) &= \int_0^\infty \frac{1}{\sigma^{n+2}} exp\left\{-\frac{z}{\sigma^2}\right\} d\sigma^2 \\
&= \int_0^\infty \left(\sigma^2\right)^{-\left(\frac{n}{2}+1\right)} exp\left\{-\frac{z}{\sigma^2}\right\} d\sigma^2 \\
&= \left(\frac{z}{2}\right)^{-\frac{n}{2}} \Gamma\left(\frac{n}{2}\right) \\
&= \left(\frac{vS^2 + \left(\beta-\hat{\beta}\right)^t X^t X \left(\beta-\hat{\beta}\right)}{2}\right)^{-\frac{n}{2}} \Gamma\left(\frac{n}{2}\right) \\
&= \Gamma\left(\frac{v+r}{2}\right) 2^{\frac{n}{2}} \left\{vS^2\left(1+\frac{\left(\beta-\hat{\beta}\right) X^t X \left(\beta-\hat{\beta}\right)}{vS^2}\right)\right\}^{-\frac{1}{2}(v+r)} \\
&\propto \Gamma\left(\frac{v+r}{2}\right) \left(vS^2\right)^{-\frac{1}{2}(v+r)} \left(1+\frac{\left(\beta-\hat{\beta}\right) X^t X \left(\beta-\hat{\beta}\right)}{vS^2}\right)^{-\frac{1}{2}(v+r)}
\end{aligned}
$$

la expresión anterior corresponde al kernel de una distribución t-Student multivariada, por lo tanto

$$
\pi\left(\beta|Y\right) = \frac{\Gamma\left(\frac{v+r}{2}\right) |X^t X|^{\frac{1}{2}} S^{-r}}{\left[\Gamma\left(\frac{1}{2}\right)\right]^r \Gamma\left(\frac{v}{2}\right) v^{\frac{r}{2}}} \left(1+\frac{\left(\beta-\hat{\beta}\right) X^t X \left(\beta-\hat{\beta}\right)}{vS^2}\right)^{-\frac{1}{2}(v+r)} \tag{4.28}
$$

es decir

$$
\pi\left(\beta|Y\right) \sim t_r\left(\hat{\beta}, S^2\left(X^t X\right)^{-1}, v\right)
$$

De lo anterior podemos concluir que para cada parámetro β_i , $i = 0, 1, ..., r$ se tiene que

$$
\pi\left(\beta_i|Y\right) \sim t\left(v, \hat{\beta}_i, h_{ii}\right)
$$

lo cual es una distribución t-Student con v grados de libertad , parámetro de posición $\hat{\beta}_i$ y parámetro de escala h_{ii} que es el elemento (i,i) de la matriz $S^2\left(X^t X\right)^{-1}$.

4.3.2. Distribución a posteriori con distribución a priori conjugada

Para encontrar la distribución en este caso supondremos que los parámetros β y σ^2 tienen una distribución a priori conocida. En el caso de β se utilizará distribución de 4.26 y para σ^2 la distribución de 4.25.

Supongamos que $\pi\left(\beta|\sigma^2, Y\right) \sim N_r\left(m, V\right)$ y $\pi\left(\sigma^2|Y\right) \sim IG\left(a, d\right)$.

Por el Teorema de Bayes sabemos que

$$\begin{aligned}
\pi\left(\beta, \sigma^2|Y\right) &\propto L(\beta, \sigma^2|Y)\pi\left(\beta, \sigma^2\right) \\
&\propto \frac{1}{\sigma^n} exp\left\{-\frac{1}{2\sigma^2}\left(Y - X\beta\right)^t\left(Y - XB\right)\right\} \\
&\quad \left(\sigma^2\right)^{-\frac{d+r+2}{2}} exp\left\{-\frac{1}{2\sigma^2}\left[\left(\beta - m\right)^t V^{-1}\left(\beta - m\right) + a\right]\right\} \\
&\propto \left(\sigma^2\right)^{-\frac{d+n+r+2}{2}} exp\left\{-\frac{1}{2\sigma^2}Q\right\}
\end{aligned}$$

con

$$\begin{aligned}
Q &= \left(Y - X\beta\right)^t\left(Y - X\beta\right) + \left[\left(\beta - m\right)^t V^{-1}\left(\beta - m\right) + a\right] \\
&= \beta^t\left(V^{-1} + X^tX\right)\beta - \beta^t\left(V^{-1}m + X^tY\right)\beta - \left(m^tV^{-1} + YX^t\right)\beta + \\
&\quad + \left(m^tV^{-1}m + Y^tY^t + a\right) \\
&= \left(\beta - m^*\right)^t\left(V^*\right)^{-1}\left(\beta - m^*\right) + a^*
\end{aligned}$$

donde

$$\begin{aligned}
V^* &= \left(V^{-1} + X^tX\right)^{-1} \\
m^* &= \left(V^{-1} + X^tX\right)^{-1}\left(V^{-1}m + X^tY\right) \\
a^* &= a + m^tV^{-1}m + Y^tY^t - \left(m^*\right)^t\left(V^*\right)^{-1}m^* \\
d^* &= d + n
\end{aligned}$$

Por lo tanto

$$\pi\left(\beta, \sigma^2|Y\right) \sim N_r\left(m^*, V^*\right) IGa\left(a^*, d^*\right)$$

Es importante mencionar que el que uso de las distribuciones iniciales elegidas fue porque así la posterior tiene la misma distribución.

Note que como X^tX es no singular

$$\begin{aligned} m^* &= \left(V^{-1}+X^tX\right)^{-1}\left(V^{-1}m+X^tX\hat{\beta}\right) \\ &= (I-A)\,m+A\hat{\beta} \end{aligned}$$

donde

$$A = \left(V^{-1}+X^tX\right)^{-1}X^tX$$

Basta observar que por un lado

$$\left(V^{-1}+X^tX\right)^{-1}V^{-1} = \left(I-X^tXV\right)^{-1}$$

y por otro

$$\begin{aligned} I-A &= I-\left(V^{-1}+X^tX\right)^{-1}X^tX \\ &= I-\left(I+X^tXV\right)^{-1}X^tXV \\ &= \left(I+X^tXV\right)^{-1}\left(I+X^tXV\right)-\left(I+X^tXV\right)^{-1}X^tXV \\ &= \left(I+X^tXV\right)^{-1}\left(I+X^tXV-X^tXV\right) \\ &= \left(I+X^tXV\right)^{-1} \end{aligned}$$

Así el parámetro m^* es una media ponderada entre la media a priori m y el estimador por mínimos cuadrados $\hat{\beta}$ con pesos respectivos $(I-A)$ y A.

Ahora supongamos que tenemos dos muestras independientes , a saber (Y_1,X_1) y (Y_2,X_2). Las distribuciones a posteriori $\pi\left(\sigma^2|Y\right)$ y $\pi\left(\beta|\sigma^2,Y\right)$ calculadas de la muestra inicial (Y_1,X_1) ,

$$\pi\left(\sigma^2\right)\sim IGa\left(\tfrac{v_1}{2},\tfrac{v_1S_1^2}{2}\right) \text{ y } \pi\left(\beta|\sigma^2\right)\sim N_r\left(\hat{\beta}_1,\hat{\sigma^2}M_0^{-1}\right)$$

con $v_1=n_1-r$, $n_1$es el tamaño de la muestra inicial , utilizando los mismos estimadores $\hat{\beta}$ y $\hat{\sigma^2}$ de la sección anterior. Por otro lado la segunda muestra de tamaño n_2, tiene la siguiente función de verosimilitud

$$L\left(\beta,\sigma^2|Y_2\right) = \frac{1}{(2\pi)^{\frac{n_2}{2}}\,\sigma^{n_2}}exp\left\{-\frac{1}{2\sigma^2}\left(Y_2-X_2\beta\right)^t\left(Y_2-X_2\beta\right)\right\}$$

Note que al tener una segunda muestra el estimador $\hat{\beta}$ tiene la siguiente forma

$$\begin{aligned}\hat{\beta} &= \left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^t \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right)^t \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \\ &= \left(\begin{bmatrix} X_1^t & X_2^t \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \right)^{-1} \begin{pmatrix} X_1^t & X_2^t \end{pmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}\end{aligned}$$

$$\hat{\beta} = \left(X_1^t X_1 + X_2^t X_2\right)^{-1} \left(X_1^t Y_1 + X_2^t Y_1\right) \tag{4.29}$$

Utilizando el Teorema de Bayes, la distribución a posteriori conjunta de β y σ^2 será

$$\begin{aligned}\pi\left(\beta, \sigma^2 | Y_2\right) &\propto L\left(\beta, \sigma^2 | Y_2\right) \pi\left(\beta, \sigma^2\right) \\ &= \frac{1}{(2\pi)^{\frac{n_2}{2}} \sigma^{n_2+r}} exp\left\{-\frac{1}{2\sigma^2}\left[v_2 S_2^2 + \left(\beta - \hat{\beta}\right)^t X_2^t X_2 \left(\beta - \hat{\beta}\right)\right]\right\} \\ &\times \left(\frac{1}{\sigma^2}\right)^{\frac{v_1}{2}+1} \\ &\times exp\left\{-\frac{1}{2\sigma^2}\left[v_1 S_1^2 + \left(\beta - \hat{\beta}\right)^t X_1^t X_1 \left(\beta - \hat{\beta}\right)\right]\right\} \end{aligned} \tag{4.30}$$

con $v_2 = n_2 - r$ y $S_2^2 = \frac{\left(Y_2 - X_2\hat{\beta}\right)^t \left(Y_2 - X_2\hat{\beta}\right)}{v_2}$.
Por la ley multiplicativa de la probabilidad , la distribución a posteriori se puede obtener de la siguiente forma

$$\pi\left(\beta, \sigma^2 | Y_2\right) = \pi\left(\sigma^2 | Y_2\right) \quad \pi\left(\beta | \sigma^2, Y_2\right) \tag{4.31}$$

con $\pi\left(\sigma^2 | Y_2\right)$ distribución marginal para σ^2 dada la segunda muestra,

$$\begin{aligned}\pi\left(\sigma^2 | Y_2\right) &\propto \frac{1}{\left(\sigma^2\right)^{\frac{n_2}{2}+\frac{v_1}{2}+1+r}} exp\left\{-\frac{v_2 S_2 + v_1 S_1}{2\sigma^2}\right\} \\ &\quad \int_{-\infty}^{\infty} exp\left\{-\tfrac{1}{2}\left(\beta - \hat{\beta}\right)^t \frac{X_2^t X_2 + X_1^t X_1}{\sigma^2}\right\}\left(\beta - \hat{\beta}\right) d\beta \\ &\propto \frac{1}{\left(\sigma^2\right)^{\frac{n_2}{2}+\frac{v_1}{2}+1+r}} exp\left\{-\frac{v_2 S_2 + v_1 S_1}{2\sigma^2}\right\} \left\{\left(\sqrt{2\pi}\right)^r \left(\frac{\left|X_2^t X_2 + X_1^t X_1\right|}{\left[\sigma^2\right]^{2r}}\right)^{-1/2}\right\} \\ &\propto \frac{1}{\left(\sigma^2\right)^{\frac{n_2}{2}+\frac{n_1-r}{2}+1+r-r}} exp\left\{-\frac{v_2 S_2 + v_1 S_1}{2\sigma^2}\right\} \\ &\propto \frac{1}{\left(\sigma^2\right)^{\frac{v}{2}+1}} exp\left\{-\frac{v S^2}{2\sigma^2}\right\}\end{aligned}$$

donde

$$v = n_1 + n_2 - r$$

$$S^2 = \frac{1}{v}\left[\left(Y_1 - X_1\hat{\beta}\right)^t\left(Y_1 - X_1\hat{\beta}\right) + \left(Y_2 - X_2\hat{\beta}\right)^t\left(Y_2 - X_2\hat{\beta}\right)\right]$$

Por lo tanto

$$\pi\left(\sigma^2|Y_2\right) \sim IGa\left(\frac{v}{2}, \frac{vS^2}{2}\right) \tag{4.32}$$

Para el segundo término de la ecuación 4.31 que es la distribución condicional a posteriori de β dados σ^2 y Y_2 obtenemos que

$$\begin{aligned}
\pi\left(\beta|\sigma^2, Y_2\right) &= \frac{\pi\left(\beta, \sigma^2|Y_2\right)}{\pi\left(\sigma^2|Y_2\right)} \\
&\propto \frac{\frac{1}{\sigma^{n_2+n_1+2}} exp\left\{-\frac{1}{2\sigma^2}\left[v_2S_2^2 + v_1S_1^2 + \left(\beta - \hat{\beta}\right)^t\left(X_2^tX_2 + X_1^tX_1\right)\left(\beta - \hat{\beta}\right)\right]\right\}}{\frac{1}{(\sigma^2)^{\frac{v}{2}+1}} exp\left\{-\frac{vS^2}{2\sigma^2}\right\}} \\
&= \frac{\left(\sigma^2\right)^{\left(\frac{n_1+n_2-r}{2}+1\right)}}{(\sigma^2)^{\left(\frac{n_1+n_2}{2}+1\right)}} \\
&\quad exp\left\{-\frac{1}{2\sigma^2}\left[v_2S_2^2 + v_1S_1^2 + \left(\beta - \hat{\beta}\right)^t\left(X_2^tX_2 + X_1^tX_1\right)\left(\beta - \hat{\beta}\right) - v_2S_2^2 - v_1S_1^2\right]\right\} \\
&\propto \sigma^{-r} exp\left\{-\frac{1}{2\sigma^2}\left(\beta - \hat{\beta}\right)^t\left(X_2^tX_2 + X_1^tX_1\right)\left(\beta - \hat{\beta}\right)\right\} \\
&= \sigma^{-r} exp\left\{-\frac{1}{2}\left(\beta - \hat{\beta}\right)^t\frac{M}{\sigma^2}\left(\beta - \hat{\beta}\right)\right\}
\end{aligned}$$

con

$$M = X_2^tX_2 + X_1^tX_1 = M_0 + X_2^tX_2$$

de esta manera la distribución condicional a posteriori tendrá una distribución normal r-variada

$$\pi\left(\beta|\sigma^2, Y_2\right) \sim N_r\left(\hat{\beta}, \sigma^2M^{-1}\right) \tag{4.33}$$

De la misma manera en que con anterioridad se obtuvo, la distribución marginal a posteriori de β resulta tener una distribución $t - Student$

$$\pi\left(\beta|Y_2\right) \quad \sim \quad t_r\left(v, \hat{\beta}, S^2 M^{-1}\right) \tag{4.34}$$

y para cada β_i se tiene que

$$\pi\left(\hat{\beta}_i|Y_2\right) \quad \sim \quad t\left(v, \hat{\beta}_i, h_{ii}\right) \tag{4.35}$$

donde h_{ii} es el elemento (i,i) de la matriz $S^2 M^{-1}$.
Finalmente se presentan las expresiones de algunos parámetros y estimadores que es importante precisar

$$v \quad = \quad n_1 + n_2 - r \tag{4.36}$$

$$M \quad = \quad X_2^t X_2 + X_1^t X_1 = M_0 + X_2^t X_2 \tag{4.37}$$

$$\hat{\beta} \quad = \quad \left(X_2^t X_2 + X_1^t X_1\right)^{-1}\left(X_1^t Y_1 + X_2^t Y_1\right) = M^{-1}\left(X_1^t Y_1 + X_2^t Y_1\right) \tag{4.38}$$

$$S^2 \quad = \quad \frac{\left(Y_1 - X_1\hat{\beta}\right)^t\left(Y_1 - X_1\hat{\beta}\right) + \left(Y_2 - X_2\hat{\beta}\right)^t\left(Y_2 - X_2\hat{\beta}\right)}{v} \tag{4.39}$$

4.3.3. Distribución predictiva

En los modelos de regresión es importante predecir los valores de una variable respuesta para futuras observaciones. Supongamos que se trata de predecir un vector Y_o de futuras observaciones, con los valores de las variables regresoras del vector X_o. En términos de los parámetros β del modelo original , las nuevas observaciones se obtienen a través de la siguiente igualdad:

$$Y_0 \quad = \quad X_0\beta + \varepsilon$$

En general , la distribución predictiva de Y se obtiene integrando sobre todo el espacio parámetral de β y σ^2. Supondremos el uso de una priori conjugada:

$$\begin{aligned} f(y) &= \int_0^\infty \int_{\mathbb{R}} L\left(\beta, \sigma^2 | Y\right) \pi\left(\beta, \sigma^2\right) d\beta d\sigma^2 \\ &\propto \int_0^\infty \int_{\mathbb{R}} \left(\sigma^2\right)^{-\frac{d+n+r+2}{2}} exp\left\{-\frac{1}{2\sigma^2} Q\right\} d\beta d\sigma^2 \end{aligned}$$

con

$$\begin{aligned} Q &= (Y - X\beta)^t (Y - X\beta) + \left[(\beta - m)^t V^{-1} (\beta - m) + a\right] \\ &= \beta^t \left(V^{-1} + X^t X\right) \beta - \beta^t \left(V^{-1} m + X^t Y\right) \beta - \left(m^t V^{-1} + Y X^t\right) \beta + \\ &\quad + \left(m^t V^{-1} m + Y^t Y^t + a\right) \\ &= (\beta - m^*)^t (V^*)^{-1} (\beta - m^*) + a^* \end{aligned}$$

donde

$$\begin{aligned} V^* &= \left(V^{-1} + X^t X\right)^{-1} \\ m^* &= \left(V^{-1} + X^t X\right)^{-1} \left(V^{-1} m + X^t Y\right) \\ a^* &= a + m^t V^{-1} m + Y^t Y^t - (m^*)^t (V^*)^{-1} m^* \\ d^* &= d + n \end{aligned}$$

Dado que

$$\pi\left(\beta, \sigma^2 | Y\right) \sim NIG\left(a^*, d^*, m^*, V^*\right)$$

El lector debe saber que la notación NIG denota la distribución normal inversa gamma y que es igual al producto de la distribución normal y la distribución gamma inversa, como se ha observado en las secciones anteriores.Recordemos que:

$$\pi\left(\beta, \sigma^2 | Y\right) = \pi\left(\beta | \sigma^2, Y\right) \pi\left(\sigma^2 | Y\right)$$

$$\Longrightarrow \pi\left(\beta | \sigma^2, Y\right) \sim N\left(m^*, \sigma^2 V^*\right)$$

$$\Longrightarrow X_0 \pi\left(\beta | \sigma^2, Y\right) \sim N\left(X_0 m^*, \sigma^2 X_0 V^* X_0^t\right)$$

Por otro lado

$$Y_0 - X_0 \beta \sim N\left(0, \sigma^2 I\right)$$

Tenemos que $(Y_0 - X_0\beta)$ y $(X_0\beta)$ son independientes, además $Y_0 = (Y_0 - X_0\beta) + (X_0\beta)$, así la distribución a predictiva a posteriori de Y_0 es

$$Y_0 \sim N\left(X_0 m^*, \sigma^2\left(I + X_0 V^* X_0^t\right)\right)$$

Como

$$\pi\left(Y_0, \sigma^2|Y\right) = \pi\left(Y_0|\sigma^2, Y\right)\pi\left(\sigma^2|Y\right)$$

De esta manera tenemos que

$$\pi\left(Y_0, \sigma^2|Y\right) \sim NIG\left(a^*, d^*, X_0 m^*, I + X_0 V^* X_0^t\right)$$

Y así la distribución marginal de Y_0 es una $t - Student\ multivariada$

$$Y_0 \sim t_{d*}\left(X_0 m^*, a^*\left(I + X_0 V^* X_0^t\right)\right)$$

4.3.4. Inferencia Bayesiana para β y σ^2

Como se ha mencionado a lo largo del presente trabajo toda la información para los parámetros se encuentra en la distribución a posteriori, deseamos hacer inferencia Bayesiana para cada parámetro: hacer estimaciones puntuales y encontrar regiones HPD.

Estimación puntual

Para hacer la estimación del parámetro β , necesitamos calcular la moda a posteriori. Sin embargo, la distribución marginal a posteriori de β es una distribución $t - multivariada$ (ver 4.9) por lo que la moda a posteriori coincide con la media y la mediana (por la simetría de la distribución), entonces el estimador de β es $\hat{\beta}$ que se muestra en la ecuación 4.38.

Por otro lado para encontrar el estimador máximo verosímil para σ^2 , debemos encontrar la moda de $\pi\left(\sigma^2|Y\right)$ que corresponde a una distribución gamma-inversa , es decir,

$$\pi\left(\sigma^2|Y\right) = \left(\frac{vS^2}{2}\right)^{\frac{v}{2}} \frac{1}{\Gamma\left(\frac{v}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{v}{2}+1} exp\left\{-\frac{vS^2}{2\sigma^2}\right\}$$

aplicando logaritmo natural a la distribución anterior tenemos

$$\ln\left(\pi\left(\sigma^2|Y\right)\right) = \frac{v}{2}\ln\left(\frac{vS^2}{2}\right) + \ln\left(\frac{1}{\Gamma\left(\frac{v}{2}\right)}\right) - \left(\frac{v}{2}+1\right)\ln\left(\sigma^2\right) - \frac{vS^2}{2\sigma^2}$$

Derivamos con respecto a σ^2 e igualando a cero

$$\frac{d\ln\left(\pi\left(\sigma^2|Y\right)\right)}{d\sigma^2} = -\left(\frac{v}{2}+1\right)\left(\sigma^2\right)^{-1}+\frac{vS^2}{2}\left(\sigma^2\right)^{-2}=0$$

despejando tenemos que

$$\sigma^2_{moda} = \frac{vS^2}{v+2} \tag{4.40}$$

Pero como la distribución gamma-inversa no es simétrica, la media a posteriori no es la misma que la moda, se elige la que tenga menor varianza. Pero conocemos la esperanza y varianza de esta distribución

$$\begin{aligned}\mu^\pi &= E^{\pi\left(\sigma^2|Y\right)}\left[\sigma^2\right]\\ &= \frac{vS^2/2}{\frac{v}{2}-1}\\ &= \frac{vS^2}{v-2}\ \ para\ \frac{v}{2}>1\end{aligned}$$

el error de estimación o varianza a posteriori para μ^π es

$$\begin{aligned}V^\pi_{\mu^\pi} &= E^{\pi\left(\sigma^2|Y\right)}\left[\left(\sigma^2-\mu^\pi\right)^2\right]\\ &= \frac{\left(\frac{vS^2}{2}\right)^2}{\left(\frac{v}{2}-1\right)^2\left(\frac{v}{2}-2\right)}\\ &= \frac{2v^2S^4}{(v-2)^2(v-4)}\ para\ \frac{v}{2}>2\end{aligned}$$

por otro lado el error de estimación de σ^2_{moda} se obtiene utilizando 3.1,

$$\begin{aligned}V^\pi_{\sigma^2_{moda}} &= V^\pi+\left(\mu^\pi-\sigma^2_{moda}\right)^2\\ &= \frac{2v^2S^4}{(v-2)^2(v-4)}+\left(\frac{vS^2}{v-2}-\frac{vS^2}{v+2}\right)\\ &= \frac{2v^2S^4}{(v-2)^2(v-4)}+\frac{16v^2S^4}{(v^2-4)^2}\end{aligned}$$

4.4. Heterocedasticidad

El modelo de regresión lineal es tal vez la herramienta estadística de más amplio uso en diversos ámbitos del conocimiento. El modelo de regresión lineal simple presenta una estructura sencilla, útil para representar gran cantidad de fenómenos reales de una manera aproximada.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

donde $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

Un problema conocido es el de la heterocedasticidad, es decir, la violación de la condición de varianza constante de los errores.

Un modelo de heterocedasticidad que se asume con frecuencia es $\sigma_i^2 = x_i^v \sigma^2$, donde v se convierte en otro parámetro del modelo, el cual desde el punto de vista clásico se estima primero y se realiza una transformación antes de aplicar los procedimientos de estimación usuales.[15]

Opuesto a la estimación del enfoque clásico, la estadística Bayesiana produce distribuciones posteriores de las cantidades desconocidas, teniendo en cuenta tanto los datos, como las distribuciones a priori sobre estos parámetros.

La función de verosimilitud será:

$$\begin{aligned} L\left(y|x, \beta_0, \beta_1, \sigma^2, v\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} x_i^{v/2} \sigma} exp\left\{-\frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{x_i^v \sigma^2}\right\} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma \prod_{i=1}^{n} x_i^{v/2}} exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{x_i^v}\right\} \end{aligned}$$

Si asumimos una distribución a priori no informativa sobre $\left(\beta, \sigma^2, v\right)$ tal que:

$$\begin{aligned} \pi\left(\beta\right) &\propto c \\ \pi\left(\sigma^2\right) &\propto \frac{1}{\sigma} \\ \pi\left(v\right) &\propto k \end{aligned}$$

donde c y k son constantes, entonces la distribución posterior de $\left(\beta, \sigma^2, v\right)$ es:

$$\pi\left(\beta, \sigma^2, v|Y\right) \propto \frac{1}{\sigma^2 \prod_{i=1}^{n} x_i^{v/2}} exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{x_i^v}\right\}$$

La anterior expresión puede aproximarse utilizando métodos de simulación conocidos como MCMC (Monte Carlo Markov Chain) utilizados para resolver problemas Bayesianos de gran complejidad (ver apéndice C). Uno de los procedimientos es el muestreo de Gibbs. En términos generales, la idea del muestreo de Gibbs es la de sobreponer las dificultades del cálculo de $\pi\left(\beta, \sigma^2, v|Y\right)$ con una sucesión de cálculos más sencillos de distribuciones condicionales. De manera muy ilustrativa se describen los principales pasos de este algoritmo:

1. Seleccionar un punto arbitrario $\left(\beta_0, \sigma_0^2, v_0\right)$
2. Generar:

- β_i de $\pi\left(\beta|Y, \sigma_{i-1}^2, v_{i-1}\right)$
- σ_i^2 de $\pi\left(\sigma^2|Y, \beta_i, v_{i-1}\right)$
- v_i de $\pi\left(v|Y, \beta_i, \sigma_i^2\right)$

Repetimos el paso anterior un número grande de veces.
Este proceso genera muestras de un proceso Markoviano donde la distribución estacionaria es la distribución posterior, descartamos los valores generados en las primeras iteraciones pues pertenecen a un periodo de aprendizaje del algoritmo.
En el paso i, (para $i \to \infty$) se obtiene lo siguiente:

$$\pi\left(\beta|Y, \sigma_{i-1}^2, v_{i-1}\right) \quad \sim \quad N\left(\hat{\beta}, \sigma_{i-1}^2\left(X^t X\right)^{-1}\right)$$

donde la $j-$*ésima* fila de X es el vector

$$\left(\frac{1}{\sigma_{i-1} x_j^{v_{i-1}/2}}, \frac{x_j}{\sigma_{i-1} x_j^{v_{i-1}/2}}\right)$$

y

$$\hat{\beta} \quad = \quad \left(X^t X\right)^{-1} X^t y^*$$

donde la $j-$*ésima* componente de y^*es

$$y_j^* \quad = \quad \frac{y_j}{\sigma_{i-1} x_j^{\frac{v_{i-1}}{2}}}$$

La distribución a posteriori condicional de σ^2 es una gamma-inversa:

$$\pi\left(\sigma^2|Y,\beta_i,v_{i-1}\right) \quad \sim \quad IGa\left(2,\frac{1}{2}\sum_{j=1}^{n}\frac{\left(y_j-(\beta_0)_i-(\beta_1)_i\,x_j\right)^2}{x_j^{v_{i-1}}}\right)$$

La distribución posterior condicional de v no tiene una forma cerrada y por lo tanto debemos utilizar algún método para la generación de valores de esta distribución (nuevamente pueden usarse el método MCMC), ya que :

$$\pi\left(v|Y,\beta_i,\sigma_i^2\right) \quad = \quad \frac{1}{\prod_{i=1}^{n}x_i^v}exp\left\{-\frac{1}{2\sigma_i^2}\sum_{j=1}^{n}\frac{\left(y_j-(\beta_0)_i-(\beta_1)_i\,x_j\right)^2}{x_j^v}\right\}$$

5. Ejemplos y discusión

5.1. Regresión Bayesiana con a priori no informativa

Supongamos que queremos hacer un ajuste Bayesiano de un modelo de regresión lineal, con errores gaussianos y una distribución inicial no informativa a los datos cuya gráfica de dispersión es la siguiente.

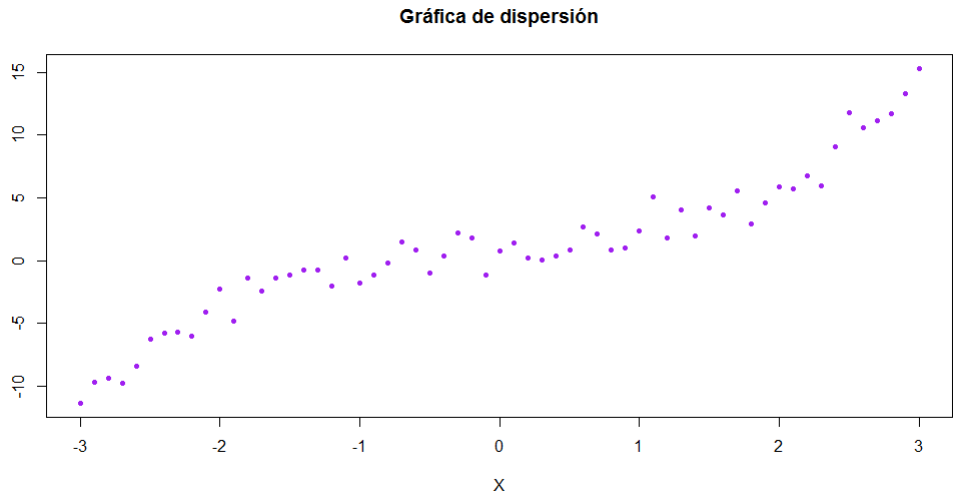


Figura 5.1: Dispersión

A partir de los datos anteriores definimos los siguientes vectores y matrices para nuestro modelo. Sean

$$X = \begin{pmatrix} 1 & -3 \\ 1 & -2.9 \\ \vdots & \vdots \\ 1 & 3 \end{pmatrix}_{61\times 2}$$

la matriz X formada por unos y las variables independientes y

$$Y = \begin{pmatrix} -11.35 \\ -9.68 \\ \vdots \\ 15.32 \end{pmatrix}$$

el vector de variables respuesta.
Así podemos obtener $\hat{\beta}$ y $\hat{y}$

$$\begin{aligned}\hat{\beta} &= \left(X^tX\right)^{-1}X^tY \\ &= \begin{pmatrix} 1.0263 \\ 3.0027 \end{pmatrix} \\ &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}\end{aligned}$$

Por lo que

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= \begin{pmatrix} -7.98 \\ \vdots \\ 10.03 \end{pmatrix}_{61\times 1}\end{aligned}$$

Asumiremos una distribución inicial $\pi\left(\beta_0, \beta_1, \sigma^2\right)$ no informativa

$$\pi\left(\beta_0, \beta_1, \sigma^2\right) \propto \frac{1}{\sigma^2}$$

De los capítulos anteriores sabemos que la distribución posterior es :

$$\pi\left(\beta_0, \beta_1, \sigma^2|Y\right) = N_2\left(\beta; \hat{\beta}, \Sigma\right) IG\left(\sigma^2; \frac{n-p-1}{2}, S^2\frac{n-p-1}{2}\right)$$

en el caso particular del ejemplo que corresponde a una regresión lineal simple

$$\begin{aligned}\Sigma &= \sigma^2\left(X^tX\right)^{-1} \\ &= \frac{\sigma^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}\end{aligned}$$

Notemos que

$$\begin{aligned}\sum x_i^2 &= 189.1 \\ n &= 61 \\ \sum x_i &= 0\end{aligned}$$

De esta manera

$$\Sigma = \sigma^2 \begin{pmatrix} 0.0163 & 0 \\ 0 & 0.00528 \end{pmatrix}$$

Adicionalmente

$$\begin{aligned} S^2 &= \frac{\left(Y-\hat{Y}\right)^t \left(Y-\hat{Y}\right)}{61-1-1} \\ &= 4.1578 \end{aligned}$$

Es decir

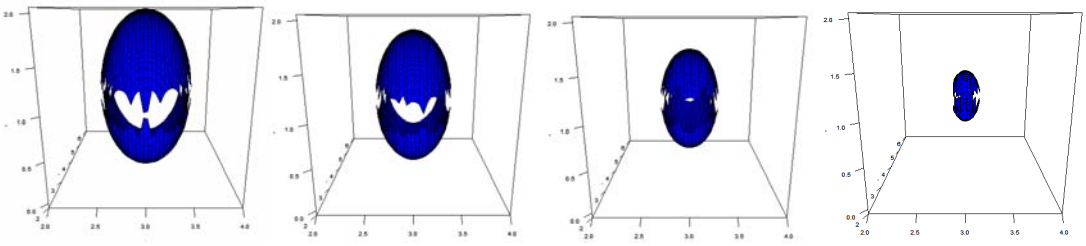
$$\begin{aligned} \pi\left(\beta_0,\beta_1,\sigma^2|Y\right) &= \frac{1}{2\pi\sqrt{det\left(\Sigma\right)}} exp\left(-\frac{1}{2}\left(\beta-\hat{\beta}\right)^t \Sigma^{-1}\left(\beta-\hat{\beta}\right)\right) \\ &\times \frac{\left(\frac{n-p-1}{2}S^2\right)^{\frac{n-p-1}{2}}}{\Gamma\left(\frac{n-p-1}{2}\right)} \left(\sigma^2\right)^{\frac{n-p-1}{2}} exp\left(-\frac{S^2\left(n-p-1\right)}{2\sigma^2}\right) \\ &= \frac{\sqrt{61\sum x_i^2-\left(\sum x_i\right)^2}}{2\pi\sigma^2} \\ &\times exp\left(-\frac{1}{2\sigma^2}\left[61\left(\beta_0-\hat{\beta_0}\right)^2+2\sum x_i\left(\beta_0-\hat{\beta_0}\right)\left(\beta_1-\hat{\beta_1}\right)+\sum x_1^2\left(\beta_1-\hat{\beta_1}\right)^2\right]\right) \\ &\times \frac{\left(29.5S^2\right)^{29.5}}{\Gamma\left(29.5\right)} \left(\sigma^2\right)^{-31.5} exp\left(-\frac{29.5}{\sigma^2}S^2\right) \end{aligned}$$

Por lo tanto

$$\begin{aligned} \pi\left(\beta_0,\beta_1,\sigma^2|Y\right) &= \frac{107.4016}{2\pi\sigma^2} exp\left(-\frac{1}{2\sigma^2}\left[61\left(\beta_0-1.026\right)^2+189.1\left(\beta_1-3.002\right)^2\right]\right) \\ &\times \frac{\left(29.5S^2\right)^{29.5}}{\Gamma\left(29.5\right)} \left(\sigma^2\right)^{-31.5} exp\left(-\frac{30S^2}{\sigma^2}\right) \end{aligned}$$

Notemos que la gráfica de $\pi\left(\beta_0,\beta_1,\sigma^2|Y\right)$ se encuentra en $\mathbb{R}^4$ y no es posible visualizarla, sin embargo podemos observar las superficies de nivel de esta distribución para distintos niveles,

$$\pi\left(\beta_0,\beta_1,\sigma^2|Y\right) = c$$



(a) c=0.001 (b) c=.01 (c) c=.1 (d) c=1

Figura 5.2: Superficies de nivel

De todo lo desarrollado en capítulos anteriores sabemos que la distribución posterior marginal para el vector β esta dada por

$$\pi\left(\beta_0, \beta_1 | y\right) \quad \sim \quad t_2\left(\beta; \hat{\beta}, \Sigma^*, 61-1-1\right)$$

donde

$$\begin{aligned} \Sigma^* &= S^2\left(X^t X\right)^{-1} \\ &= 4.157\begin{pmatrix} .0163 & 0 \\ 0 & .00528 \end{pmatrix} \\ &= \begin{pmatrix} .0681 & 0 \\ 0 & .0219 \end{pmatrix} \end{aligned}$$

De esta manera

$$\begin{aligned} \pi\left(\beta_0, \beta_1 | y\right) &= \frac{1}{2\pi\sqrt{det\left(\Sigma\right)}\left[1+\frac{1}{59}\left(\beta-\hat{\beta}\right)^t \sum^{*-1}\left(\beta-\hat{\beta}\right)\right]^{\frac{61}{2}}} \\ &= \frac{1}{2\pi\sqrt{.0014}\left(1+\frac{1}{59S^2}\left[n\left(\beta_0-\hat{\beta}_0\right)^2+\sum x_i^2\left(\beta_1-\hat{\beta}_1\right)^2\right]\right)^{\frac{61}{2}}} \\ &= \frac{1}{0.0774\pi\left(1+\frac{1}{245.3149}\left[61\left(\beta_0-1.026\right)^2+189.1\left(\beta_1-3.002\right)^2\right]\right)^{\frac{61}{2}}} \end{aligned}$$

Cuya gráfica es la siguiente

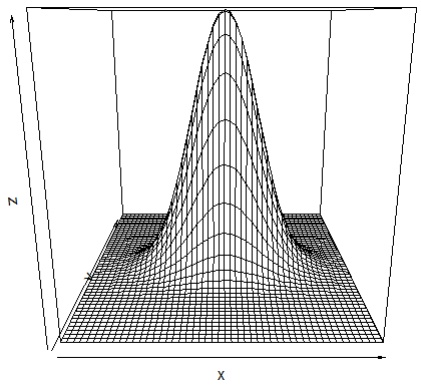


Figura 5.3: Distribución marginal del vector β

Bajo un pérdida cuadrática, el estimador Bayesiano es la esperanza condicional.Esto es

$$\begin{aligned} E[\beta|Y] &= E\left[E\left[\beta|\sigma^2\right]|Y\right] \\ &= E\left[\hat{\beta}|Y\right] \\ &= \hat{\beta} \end{aligned}$$

Por lo tanto , el estimador Bayesiano bajo una pérdida cuadrática de β es :

$$\begin{aligned} d^{\pi} &= \hat{\beta} \\ &= \begin{pmatrix} 1.0263 \\ 3.0027 \end{pmatrix} \end{aligned}$$

De los capítulos anteriores sabemos que $\sigma^2|Y \sim IG\left(\frac{n-p-1}{2}, S^2\frac{n-p-1}{2}\right)$. Por lo tanto el estimador Bayesiano bajo una pérdida cuadrática para σ^2 es :

$$\begin{aligned} d^{\pi} &= E\left[\sigma^2|Y\right] \\ &= \frac{S^2\left(\frac{n-p-1}{2}\right)}{\frac{n-p-1}{2}-1} \\ &= S^2\left(\frac{n-p-1}{n-p-3}\right) \\ &= 153.45 \end{aligned}$$

Sabemos en general que una manera equivalente de describir la región de credibilidad con densidad posterior máxima al nivel de 95 % para el vector β esta dada por la elipse:

$$\{(\beta_0, \beta_1) \quad | \left(\beta - \hat{\beta}\right)^t X^t X \quad \left(\beta - \hat{\beta}\right) \leq 2S^2 F_{2,59}(0.05)\}$$

Donde :

- $F_{2,59}(0.05)$ es el cuantil .95 de una distribución $F_{2,59}$. En este caso $F_{2,59}(0.05) = 3.1531$
-

$$\begin{aligned}\left(\beta - \hat{\beta}\right)^t X^t X \left(\beta - \hat{\beta}\right) &= n\left(\beta_0 - \hat{\beta}_0\right)^2 + 2\sum x_i \left(\beta_0 - \hat{\beta}_0\right)\left(\beta_1 - \hat{\beta}_1\right) + \sum x_i^2 \left(\beta_1 - \hat{\beta}_1\right)^2 \\ &= 61\left(\beta_0 - 1.026\right)^2 + 189\left(\beta_1 - 3.002\right)^2\end{aligned}$$

Por lo tanto la región de credibilidad con densidad posterior máxima al nivel de 95 % para el vector β es:

$$\{(\beta_0, \beta_1) \quad | \quad 61\left(\beta_0 - 1.026\right)^2 + 189\left(\beta_1 - 3.002\right)^2 \leq 26.22\}$$

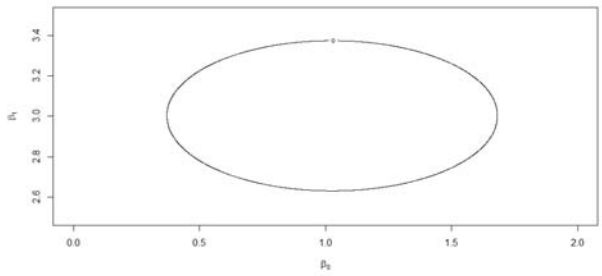


Figura 5.4: Región HPD para el vector β

Ahora encontraremos los intervalos HPD para β_0 y β_1, y recordamos que:

$$\frac{\beta_j - \hat{\beta}_j}{\sqrt{S^2 d_{jj}}} | Y \quad \sim \quad t_{59}, para\ j = 0, 1$$

donde d_{jj} es la entrada (j, j) de la matriz :

$$\begin{aligned}\left(X^t X\right)^{-1} &= \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \\ &= \begin{pmatrix} 0.0163 & 0 \\ 0 & .0052 \end{pmatrix}\end{aligned}$$

Entonces, el intervalo HPD al 95 % para β_j esta dado por:

$$(\hat{\beta}_j - \sqrt{S^2 d_{jj}} t_{59} \quad , \hat{\beta}_j + \sqrt{S^2 d_{jj}} t_{59} \quad)$$

con $t_{59} = 2.0009$

Por los tanto, los intervalos HPD serán:

Para β_0:	$(0.5039, 1.5488)$
Para β_1:	$(2.7060, 3.2994)$

Finalmente utilizando los estimadores Bayesianos bajo una pérdida cuadrática , en la siguiente gráfica se muestra el ajuste y sus intervalos de densidad posterior al 95 %.

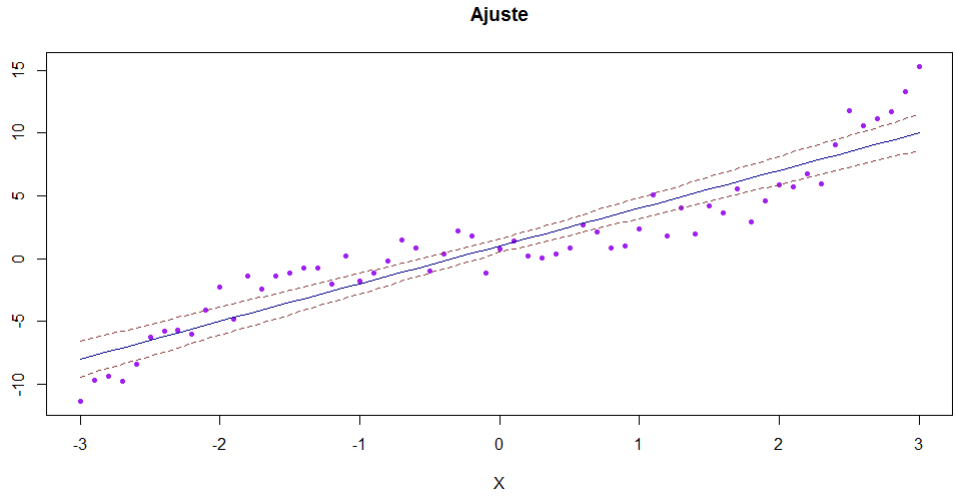


Figura 5.5: Ajuste

Es importante recalcar que los datos utilizados en presente ejemplo claramente no son lineales. Se presentó el ajuste de un modelo lineal a estos para ilustrar la metodología Bayesiana. A continuación se presenta la gráfica de dispersión de otros datos (que a simple vista parecen tener una relación lineal) y su ajuste Bayesiano.

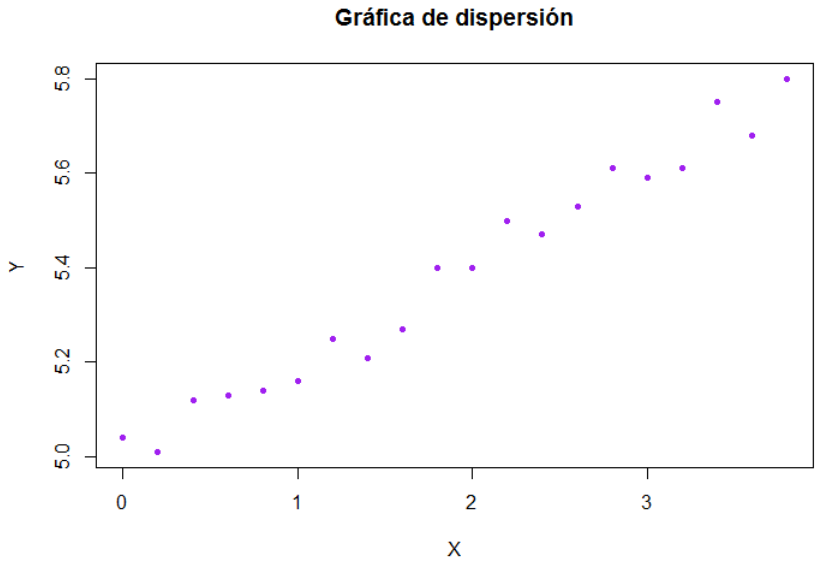


Figura 5.6: Dispersión

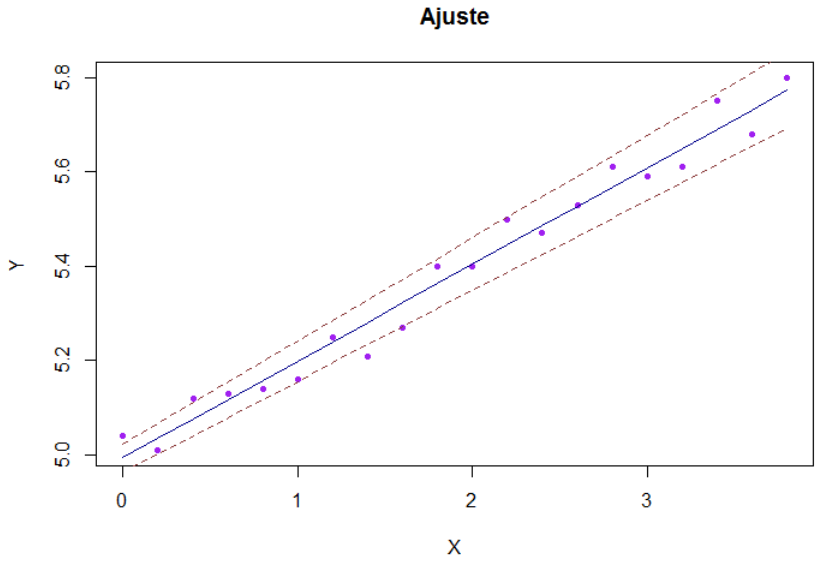


Figura 5.7: Ajuste Bayesiano.

En general podemos decir que si se hace el ajuste de un modelo lineal (ya sea con el enfoque clásico o Bayesiano) a datos que no son lineales una recta nunca será la mejor manera de explicar la relación entre estos. Sin embargo, en otros contextos el modelo lineal es de gran ayuda si se ajusta a datos cuya relación no es lineal, por ejemplo, para ayudar a estimar la tendencia en una serie de tiempo.

5.2. Regresión Bayesiana con a priori conjugada

Los siguientes datos corresponde a la temperatura promedio (medida en grados Celsius) y la elevación de ciertos puntos en el noroeste de México para el mes de enero de 2014 (estos datos fueron obtenidos de INEGI). A continuación se muestra el histograma para la temperatura promedio.

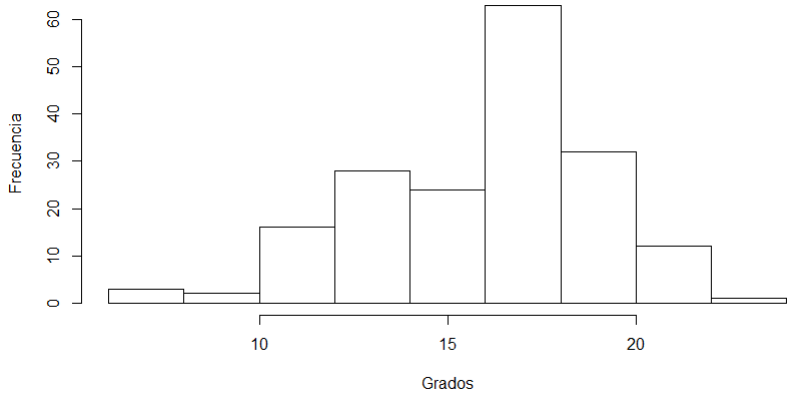


Figura 5.8: Histograma Temperatura Promedio

Deseamos explicar a través de un modelo de regresión lineal Bayesiano la relación entre la temperatura promedio y la elevación. La gráfica de dispersión de estas dos variables es la siguiente:

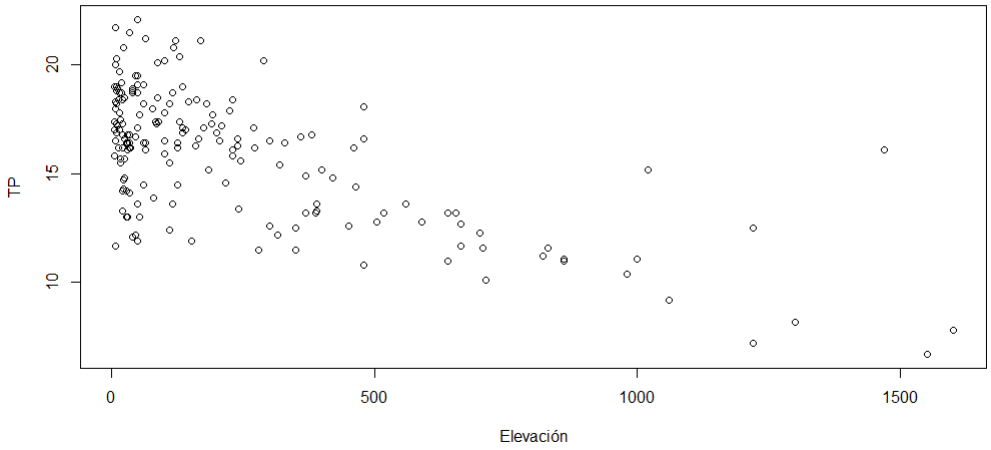


Figura 5.9: Dispersión

A partir de los datos anteriores definimos los siguientes vectores y matrices para el modelo. Sea

$$X = \begin{pmatrix} 1 & 130 \\ 1 & 230 \\ \vdots & \vdots \\ 1 & 12 \end{pmatrix}_{181\times 2}$$

la matriz X formada por las variables independientes y por otro lado

$$Y = \begin{pmatrix} 20.4 \\ 18.4 \\ \vdots \\ 17.2 \end{pmatrix}$$

el vector de variables respuesta.

Para este ejemplo utilizaremos una distribución a priori conjugada , por lo que supondremos los siguientes parámetros para las distribuciones iniciales:

$$\begin{aligned} \pi(\beta) &\sim N(0,9) \\ \pi\left(\sigma^{-2}\right) &\sim Gamma(4,6) \end{aligned}$$

En este ejemplo haremos uso de la función *MCMCregress* que se encuentra en la librería *MCMCpack* del sofware R project.

Esta función genera una muestra de la distribución posterior de un modelo de regresión lineal con errores Gaussianos utilizando el muestreo de Gibbs (con una distribución gaussiana multivariante para $\beta|Y,\sigma^2$ y gamma inversa para $\sigma^2|Y$). Utilizando el muestreo de Gibbs con $10,000$ iteraciones (utilizando $1,000$ iteraciones de calentamiento) obtenemos los siguientes resultados:

La gráficas siguientes muestran los procesos Markovianos para estimar las densidades posteriores de $\beta|Y,\sigma^2$ y $\sigma^2|Y$, las densidades posteriores y los promedios ergódicos.

Los promedios ergódicos estan definidos de la siguiente manera:

$$g\left(X_i\right) = \frac{\sum_{j=1}^{i} X_j}{j}$$

donde $\{X_i\}_{i\epsilon I}$ es una Cadena de Markov.

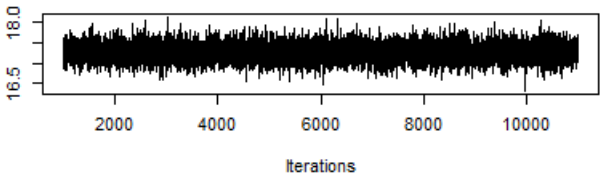


Figura 5.10: Proceso Markoviano para β_0.

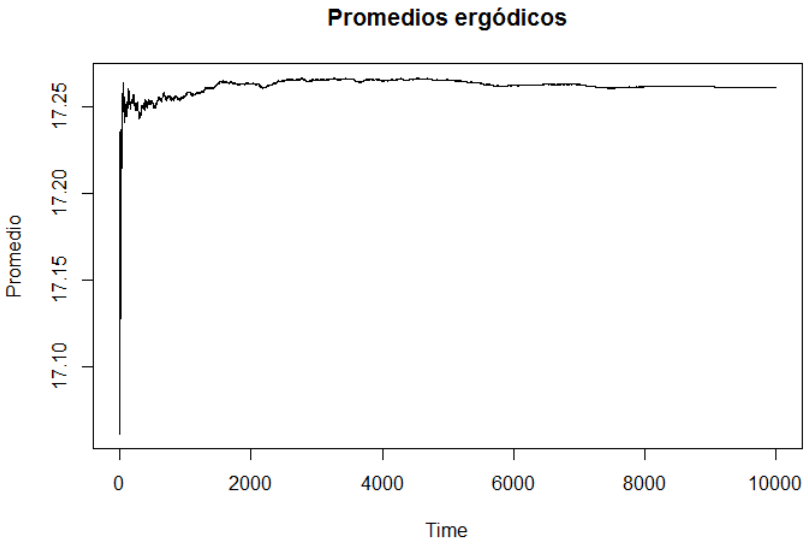


Figura 5.11: Promedios ergódicos para β_0.

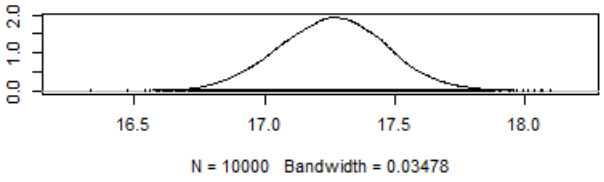


Figura 5.12: Densidad posterior para β_0.

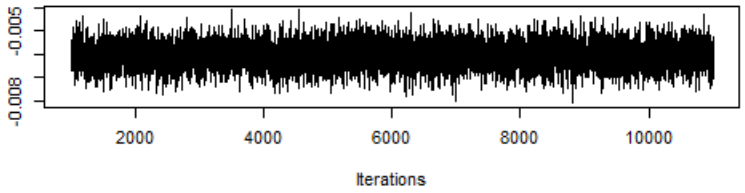


Figura 5.13: Proceso Markoviano para β_1.

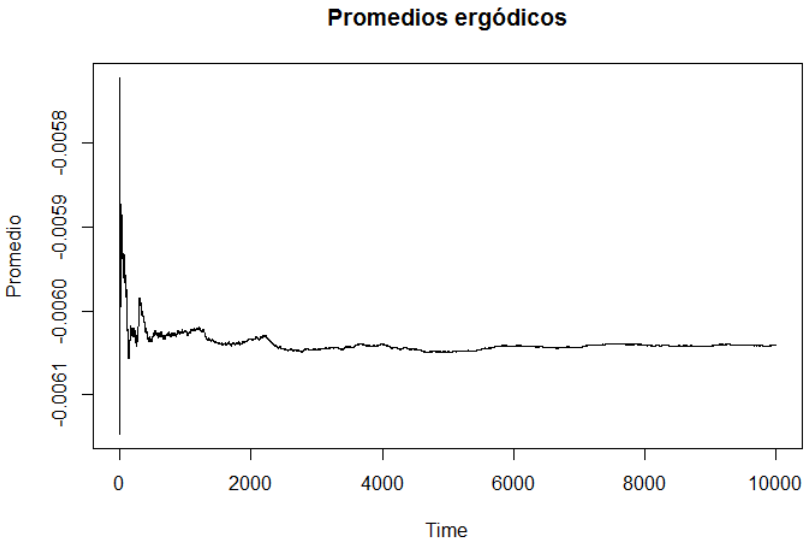


Figura 5.14: Promedios ergódicos para β_1.

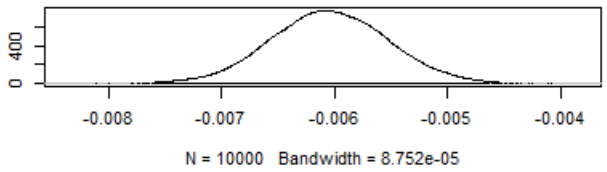


Figura 5.15: Densidad posterior para β_1.

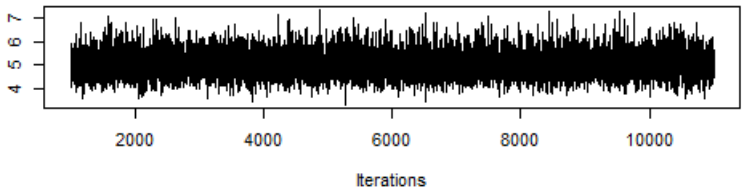


Figura 5.16: Proceso Markoviano para σ^2.

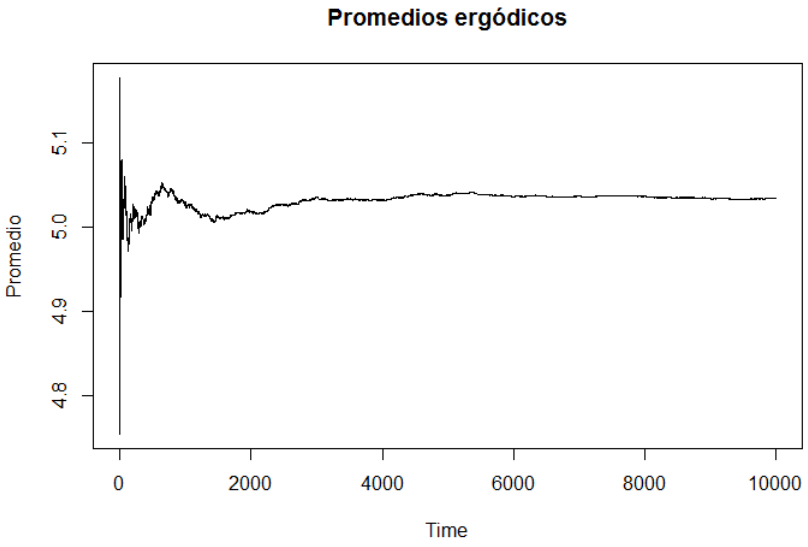


Figura 5.17: Promedios ergódicos para σ^2.

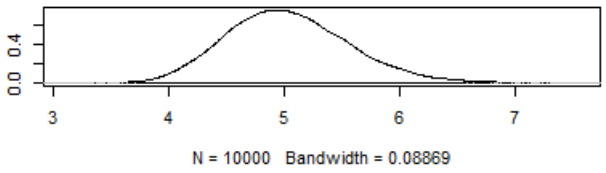


Figura 5.18: Densidad posterior σ^2.

Para nuestro ejemplo el periodo de calentamiento $(1,000)$ es el adecuado pues para los tres casos se cumple que

$$|g(X_i) - g(X_{i+1})| \quad < \quad \delta$$

para toda $\delta > 0$ e $i > 1,000$.

Cada una de las Cadenas de Markov gráficadas anteriormente converge al valor del parámetro de interés. La siguiente tabla muestra el parámetro y el valor al cuál converge su respectiva cadena:

Parámetro	Valor
β_0	17.260820
β_1	-0.006042
σ^2	5.034929

Por lo que la recta de regresión queda dada por:

$$Y = X \begin{pmatrix} 17.260820 \\ -0.006042 \end{pmatrix}$$

En la figura siguiente se puede observar el ajuste de la recta a los datos.

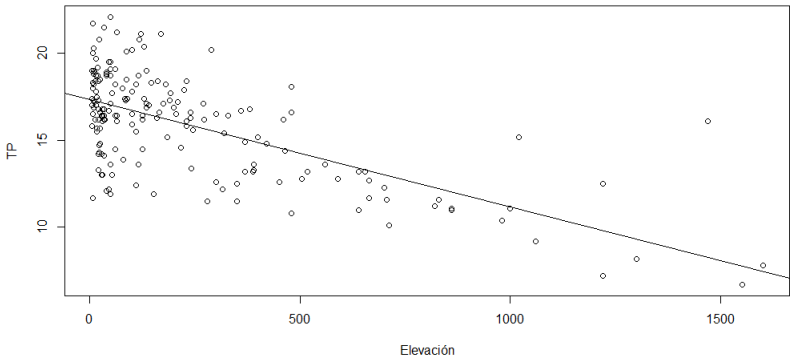


Figura 5.19: Ajuste

6. Conclusiones

El enfoque estadístico usado la mayoría de las veces es el enfoque frecuentista, este enfoque es el que se desarrolla a partir de los conceptos de probabilidad y que se centra en el cálculo de probabilidades y en los contrastes de hipótesis. De cierta manera, la estadística clásica tiene por objetivo determinar una conclusión, ya sea con la significancia estadística o aceptación y rechazo de hipótesis, siempre en el marco del estudio que se esté realizando. Por ejemplo, en un análisis estadístico que pretende comparar la eficacia de un nuevo tratamiento frente a otro, únicamente se utiliza la información obtenida de la muestra, es decir, los datos. De esta manera no existen subjetividades referentes a los parámetros, ya que se han fijado los criterios de decisión desde un principio y estos permanecen estáticos durante todo el estudio.

En un enfoque alternativo a la estadística frecuentista, aparece en escena cada vez más el enfoque Bayesiano, basada y como su nombre lo indica en el teorema de Bayes, y que se diferencia del enfoque clásico básicamente en la incorporación de información externa al estudio que se esta realizando. Si se conoce la probabilidad de que ocurra un suceso, su valor será cambiado cuando se disponga de esa información. De esta manera, las fuentes de información *a priori* se ven transformadas en probabilidad *a posteriori* y se utiliza para realizar la inferencia.

La regresión Bayesiana es un modelo muy sensitivo, pues al formar parte de una estadística subjetiva contenida en un ambiente de probabilidades, se podrían presentar varios problemas al momento de hacer un análisis. Al elegir una distribución *a priori* u otra de los coeficientes de regresión se pueden cometer errores al momento de tomar estas decisiones como investigador y llegar a un modelo que no explique del todo los datos. Por lo que la regresión Bayesiana depende de que se tenga una buena información *a priori* para así obtener una distribución posterior efectiva.

Dentro del análisis estadístico, la metodología Bayesiana, comienza resumiendo cuantitativamente la información previa existente de origen diverso . Por ejemplo, datos de laboratorio a través de experimentos, estudios publicados e incluso opinión de expertos. A partir de esto, se amplía a otro contexto que hace referencia a la definición subjetiva de probabilidad, es decir, a la convicción personal del investigador para emitir juicios acerca de una hipótesis. Sin embargo , el método frecuentista no está exento de subjetividad, puesto que el nivel de significancia, así como las explicaciones de las causalidad que se concede a determinados resultados apuntan inevitablemente hacia la apreciación propia de cada investigador.

El área de aplicación de la metodología Bayesiana es la misma que la del enfoque clásico, pero ofrece mayores ventajas en algunas situaciones. Sencillamente son dos métodos para el análisis estadístico.

En estudios actuales se aplican tanto el enfoque clásico como el Bayesiano, aunque en muchas ocasiones, el razonamiento coincide más con el Bayesiano al

tratar de utilizar la información que se conoce por parte del investigador o la experiencia.

Posiblemente, no sea muy importante elegir un método u otro como herramienta de análisis. Sin embargo el hecho de poder analizar los resultados y entender lo que se esta haciendo, en contraposición a una aplicación casi mecánica del enfoque clásico, dificultan la elección del enfoque Bayesiano.

Apéndice

A. Distribuciones de probabilidad

Recordamos en esta sección la densidad y los dos primeros momentos de la mayoría de las distribuciones utilizadas en esta tesis.

A.1. Distribución Normal, $N_p(\theta, \Sigma)$

$\theta \in \mathbb{R}^p$ y Σ es una matriz de orden $p \times p$ definida positiva y simétrica

$$f(x|\theta,\Sigma) = (det\Sigma)^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} exp\left\{-\frac{(x-\theta)^t \left(\Sigma^{-1}\right)(x-\theta)}{2}\right\}$$

$$\begin{aligned} E[x] &= \theta \\ E\left[(x-\theta)(x-\theta)^t\right] &= \Sigma \end{aligned}$$

Cuando Σ no está definida, la distribución $N_p(\theta, \Sigma)$ no tiene densidad con respecto a la medida de Lebesgue sobre $\mathbb{R}^p$. Para $p = 1$, la distribución *log-normal* está definida como la distribución de e^X cuando $X \sim N\left(\theta, \sigma^2\right)$.

A.2. Distribución Gamma, $Gamma(\alpha, \beta)$

$\alpha,\ \beta > 0$

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} I_{[0,\infty)}(x)$$

$$\begin{aligned} E[x] &= \frac{\alpha}{\beta} \\ Var[x] &= \frac{\alpha}{\beta^2} \end{aligned}$$

Los casos particulares de la distribución gamma son la distribución de Erlang que corresponde a una $Gamma(\alpha, 1)$, la distribución exponencial que es una $Gamma(1, \beta)$ (denotada por $Exp(\beta)$) y la distribución ji-cuadrada $Gamma(v/2, 1/2)$ (denotada por χ^2_v).

A.3. Distribución t de Student, $t_p(v, \theta, \Sigma)$

$v > 0$, $\theta \in \mathbb{R}^p$ y Σ es una matriz de orden $p \times p$ definida positiva y simétrica

$$f(x, v, \theta, \Sigma) = \frac{\Gamma\left(\frac{v+p}{2}\right)/\Gamma\left(\frac{v}{2}\right)}{(det\Sigma)^{\frac{1}{2}}(v\pi)^{\frac{p}{2}}}\left[1 + \frac{(x-\theta)^t\left(\Sigma^{-1}\right)(x-\theta)}{v}\right]^{-\frac{v+p}{2}}$$

$$\begin{aligned} E[x] &= \theta,\ v > 1 \\ E\left[(x-\theta)(x-\theta)^t\right] &= \frac{v\Sigma}{v-2},\ v > 2 \end{aligned}$$

A.4. Distribución Gamma Inversa, $IGa(\alpha, \beta)$

α, $\beta > 0$

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{e^{-\beta/x}}{x^{\alpha+1}}$$

$$\begin{aligned} E[x] &= \frac{\beta}{\alpha-1}\ \alpha > 1 \\ Var[x] &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}\ \alpha > 2 \end{aligned}$$

B. Teoremas

B.1. Teoremas relacionados con matrices

Teorema B.1. *Sea A una matriz idempotente. Entonces:*

$$rango(A) = tr(A)$$

Teorema B.2. *Sea $B \in M_{q\times n}(\mathbb{R})$. Entonces:*

1. *Si $rango(B) = q$, entonces B^2 es positiva definida.*
2. *Si $rango(B) < q$, entonces B^2 es positiva semidefinida.*

B.2. Teoremas de variables aleatorias

Teorema B.3. *Sea A una matriz cuadrada de tamaño $k \times k$ y Y un vector aleatorio normal multivariado de tamaño k, con media μ y matriz no singular de covarianza $\sigma^2 I$ $\left(Y \sim N_k\left(\mu, \sigma^2 I\right)\right)$. Sea U la forma cuadrática dada por $U = Y^t A Y$. Si A es idempotente y de rango h, entonces*

$$\frac{U}{\sigma^2} \sim \chi^2_{h,\lambda}$$

con $\lambda = \frac{\mu^t A \mu}{\sigma^2}$
Es decir, tiene distribución χ^2 no centrada con h grados de libertad y parámetro de no centralidad λ.

C. 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. Desafortunadamente, en la práctica la distribución posterior puede ser tan compleja que, generalmente, las integrales que implican, pueden no resolverse de manera simple, por lo cual es necesario hacer uso 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), que son empleados para distribuciones posteriores unimodales.
- Métodos Monte Carlo no iterativos (muestreo directo y muestreo indirecto).
- Métodos MCMC (Metrópolis-Hasting y Muestreo 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

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.[9]

C.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 queremos encontrar:

$$I = \int_a^b f(y)\, dy$$

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

$$\int_a^b f(y)\, dy = \int_a^b f(y) \frac{g(y)}{g(y)} dy$$

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

$$I = \int_a^b f(y) \frac{g(y)}{g(y)} dy = \int_a^b h(y)\, g(y)\, dy = E(h(y))$$

Para obtener este valor esperado es posible simular variables aleatorias independientes e idénticamente distribuidas $Y_1, ..., Y_n$, con función de densidad $g(y)$. Por la ley fuerte de los grandes números, se tiene

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

Por lo tanto

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

es un estimador consistente e insesgado de I.

C.2. Métodos Monte Carlo 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. En términos probabilísticos lo anterior queda expresado de la siguiente manera:

$$P[\theta_{n+1} \in S | \theta_n \in S_n, \theta_{n-1} \in S_{n-1}, ..., \theta_0 \in S_0] = P[\theta_{n+1} \in S | \theta_n \in S_n]$$

para todos los conjuntos $S_0, ..., S_n \subset \sigma(S)$, donde S denota el espacio de estados de la cadena y $\sigma(S)$ es una $\sigma - \acute{a}lgebra$ 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:

- Homogénea, es decir, que $P[\theta_{n+1} \in S | \theta_n \in S_n]$ no dependa de n.
- Irreducible, desde cualquier estado se puede acceder a otro, todos los estados se comunican entre sí.
- Aperiódica.
- Reversible.

Es importante mencionar que la propiedad de reversibilidad asegura la existencia de una cadena ergódica. Cuando el espacio de estados S 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 S es no numerable, que es el caso más común en aplicaciones del algoritmo de Metropolis-Hastings y Muestreo de Gibbs, existen otras consideraciones adicionales.

Sea $\{\theta_t\}_{t\in T}$ una cadena de Markov homogénea, irreducible y aperiodica, con espacio de estados S y distribución de equilibrio $\pi(\theta)$. Entonces, conforme $t \to \infty$:

1. $\theta_t \to \theta$, donde $\theta \sim \pi(\theta)$

2. $\frac{1}{t}\sum_{i=1}^{t} g(\theta_i) \rightarrow E[g(\theta)]$, casi seguramente

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

C.3. Metropolis-Hastings

Este algoritmo permite construir una cadena de Markov, los pasos a seguir se describen a continuación.

Sea $q(\theta^*, \theta)$ una distribución de transición arbitraria y definimos

$$\alpha(\theta^*, \theta) = min\left\{\frac{\pi(\theta^*) q(\theta, \theta^*)}{\pi(\theta) q(\theta^*, \theta)}, 1\right\}$$

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 $q(., \theta)$. Un nuevo valor θ^* es aceptado con probabilidad α. La cadena puede permanecer en un mismo estado a lo largo de muchas iteraciones. En la práctica, se acostumbra monitorear esto para calcular el porcentaje medio de iteraciones para los cuales los nuevos valores se aceptan.

Dado un valor inicial θ_0, la $t - ésima$ iteración del algoritmo consiste en:

1. Generar una observación θ^* de $q(\theta^*, \theta)$.
2. Generar una variable aleatoria $u \sim U(0, 1)$.
3. Si $u \leq \alpha(\theta^*, \theta_t)$ hacer $\theta_{t+1} = \theta^*$ en caso contrario $\theta_{t+1} = \theta_t$.

De esta manera se genera una cadena de Markov con probabilidad de transición a un paso dada por:

$$P(\theta_t, \theta_{t+1}) = \alpha(\theta_{t+1}, \theta_t) q(\theta_{t+1}, \theta_t)$$

La probabilidad de aceptación $\alpha(\theta^*, \theta)$ sólo depende de $\pi(\theta)$ 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 $q(\theta^*, \theta) = q(\theta^*, \theta)$, es decir, una densidad simétrica centrada en el origen. entonces:

$$\alpha(\theta^*, \theta) = min\left\{\frac{\pi(\theta^*)}{\pi(\theta)}, 1\right\}$$

- 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 $\omega(\cdot) = \pi(\cdot) q_0(\cdot)$.

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.

Para el caso de caminata aleatoria se suele considerar

$$q(\theta^*, \theta) \sim N_d\left(\theta, kV\left(\hat{\theta}\right)\right)$$

y para el caso independiente

$$q_0(\theta^*) \sim N_d\left(\hat{\theta}, kV\left(\hat{\theta}\right)\right)$$

donde $\hat{\theta}$ y $V\left(\hat{\theta}\right)$ denotan a la media y a la matriz de varianzas-covarianzas de la aproximación asintótica normal para $\pi(\theta)$, respectivamente, y $k \geq 1$ es un factor de sobredispersión, porque se ajusta mejor a las características de los datos.

C.4. Algoritmo de Gibbs

El algoritmo de Gibbs permite simular una cadena de Markov $\{\theta_t\}_{t \in T}$ con distribución de equilibrio $\pi(\theta)$. Cada valor nuevo de la cadena se obtiene al generar muestras de distribuciones cuya dimensión es menor y que en la mayoría de los casos tiene una forma más sencilla que la de $\pi(\theta)$.

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

$$\begin{gathered}
\pi(\theta_1|\theta_2, ..., \theta_p) \\
\vdots \\
\pi(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p) \ \ \forall i = 2, ..., p-1 \\
\vdots \\
\pi(\theta_n|\theta_1, ..., \theta_{n-1})
\end{gathered}$$

Estas densidades pueden identificarse al analizar la forma de la distribución final $\pi(\theta)$.De hecho para cada $i = 1, 2, ..., p$:

$$\pi(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p) \quad \propto \quad \pi(\theta)$$

donde $\pi(\theta)$ es vista sólo como función de θ_i.
Dado un valor inicial $\theta^{(0)} = \left(\theta_1^{(0)}, ..., \theta_p^{(0)}\right)$, 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 $\pi\left(\theta_1|\theta_2^{(t)}, ..., \theta_p^{(t)}\right)$
- generar una observación $\theta_2^{(t+1)}$ de $\pi\left(\theta_2|\theta_1^{(t+1)}, ..., \theta_p^{(t)}\right)$
- generar una observación $\theta_3^{(t+1)}$ de $\pi\left(\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, ..., \theta_p^{(t)}\right)$...
- generar una observación $\theta_p^{(t+1)}$ de $\pi\left(\theta_n|\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{p-1}^{(t+1)}\right)$

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

$$P\left(\theta^{(t)}, \theta^{(t+1)}\right) \quad = \quad \prod_{i=1}^{p} \pi\left(\theta_i^{(t+1)}|\theta_1^{(t+1)}, ..., \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t+1)}, ..., \theta_p^{(t+1)}\right)$$

Referencias

[1] Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis. Second edition. New York: Springer Verlag.

[2] Bernardo, J.M. (1981) Bioestadstica, una Perspectiva Bayesiana. Barcelona: Vicens Vives.

[3] Bernardo, J.M. & Smith, A.F.M. (1994). Bayesian Theory. Chichester: Wiley.

[4] Box, G.E.P. & Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Reading: Addison Wesley.

[5] Casella, G. & Berger, R.L. (2001). Statistical Inference. Belmont: Duxbury Press.

[6] Congdon, P. (2001). Bayesian Statistical Modelling. Chichester: Wiley.

[7] De Groot, M.H. (1970). Optimal Statistical Decisions. New York: McGraw-Hill.

[8] De Groot, M.H. (1988). Probabilidad y Estadstica. Mexico: AddisonWesley Iberoamericana.

[9] Gamerman D. & Lopes, H.F. (2006). Markov Chain Montecarlo. Stochastic Simulation for Bayesian Inference. Second edition. London: Chapman & Hall.

[10] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). Bayesian Data Analysis. Second edition. London: Chapman & Hall.

[11] Lindley, D.V. (1965). An Introduction to Probability and Statistics from a Bayesian Viewpoint. Vol 2. Inference. Cambridge: Cambridge University Press.

[12] Lindley, D.V. (1985). Making Decisions. Second edition. London: Wiley.

[13] Mignon, H.S. and Gamerman, D. (1999). Statistical Inference: An Integrated Approach. London: Arnold.

[14] O'Hagan, A. (1994). Kendalls Advanced Theory of Statistics. Vol 2b. Bayesian Inference. Cambridge: Edward Arnold.

[15] Press, S.J. (1989). Bayesian Statistics. Principles, Models and Applications. New York: Wiley.

[16] Robert, C.P. (2001). The Bayesian Choice. Second edition. New York: Springer Verlag.