Una propuesta bayesiana para realizar inferencias en poblaciones Heterocedásticas con observaciones correlacionadas

Natalia Henríquez C.*

Resumen

Es usual realizar inferencia con observaciones no correlacionadas, no obstante, en la vida real podemos encontrar datos que al parecer no poseen esta característica, por ejemplo, datos sísmicos, de esta manera es necesario investigar que procedimientos estadísticos pudiesen ayudar en esta idea y determinar si existe algún tipo de correlación significativa o no entre dichas observaciones.

El presente artículo, aborda esta idea desde el punto de vista del enfoque Bayesiano, dando un procedimiento para establecer la existencia de correlación entre observaciones de subpoblaciones heterocedásticas, las cuales entre ellas están no correlacionadas. Además este nos permite trabajar con una sola muestra aleatoria sin la necesidad de tener réplicas de las mediciones o tener tamaños muéstrales grandes para realizar el análisis inferencial, que es lo usual cuando se trabaja con análisis de varianzas o modelos de efectos aleatorios.

Palabras Claves: Enfoque Bayesiano, independencia, correlación, simulación.

A bayesian proposal for making inferences in Heteroscedastic populations with correlated observations

Abstract

Making an inference with correlated observations is very common. However, in real life we can find data which do not match this characteristic, e.g., seismic data. Thus, it is necessary to look into what statistical procedures may help make inferences and determine whether there is some kind of significant correlation between those observations.

This article addresses this idea from a Bayesian approach and provides a procedure to establish the existence of correlation between observations of heteroscedastic subpopulations, which are not correlated between them. Furthermore, this approach allows us to work with a single random sample without requiring measurement replications or larger sample sizes to make inferential analysis –usually made when using variance analysis and random effect models.

Key words: Bayesian approach, independence, correlation and simulation.

* Doctor en Estadística. Docente en Facultad de Ingeniería, Universidad Ucinf.

Introducción

La motivación principal de este artículo es presentar una propuesta para realizar inferencia en súper poblaciones finitas o infinitas, no existiendo correlación entre ellas, y bajo la característica de poseer una estructura de correlación al interior de estas.

Por ejemplo, datos como mediciones sísmicas se consideran que están correlacionados por la zona (subpoblación), y datos relacionados con el procesamiento de imágenes en los cuales la gama de colores y el color suponen una incidencia en la calidad de la imagen hacen suponer una correlación entre las mediciones de estos. Este tipo de datos por sus características es natural trabajarlos como modelos jerárquicos o modelos en multiniveles.

Los métodos más utilizados en el análisis de variables numéricas continuas están en su mayoría diseñados en los que se registra una única medida por cada unidad de observación, por ejemplo en análisis de varianza y regresión. Es poco usual poseer réplicas de mediciones de un mismo individuo y además la misma cantidad de ellas para todos los individuos observados, que es el caso cuando se trabaja con datos longitudinales (temporales) (Cannon, 2001)[8] o también en el caso de diseños balanceados (Cnaan,1997)[9]; no obstante podemos visualizar una cierta correlación entre la mediciones del individuo, por lo cual no pueden considerarse como observaciones independientes, supuesto básico para estimar un modelo de regresión clásico. Otra clase de estudios con observaciones correlacionadas son los datos agrupados, en los que existe un diseño jerárquico, por ejemplo datos agrupados en hospitales, puede ocurrir que los pacientes que se atienden en las distintas unidades tengan ciertas características muy similares entre ellos, es decir nuevamente bajo ciertos factores las observaciones podrían no ser independientes entre ellas (Leary, 2000)[16].

También estudios de crecimiento, en los que el sujeto se evalúa en diferentes edades o momentos, se consideran como investigaciones de datos correlacionados.

Modelos para el análisis de respuestas categóricas multivariadas correlacionadas han sido utilizados para medir la calidad de la docencia (Casero, 2010)[10], en los cuales se visualizan multiniveles (al menos tres), este tipo de mediciones podrían estar correlacionadas de acuerdo a la apreciación que los estudiantes poseen a priori del profesor, y esta influir en la evaluación a posteriori del docente.

Además autores (Ita *et al.*, 1998)[14], hacen una introducción sobre el modelamiento con multiniveles en presencia de datos (encuestas) correlacionados y no correlacionados (Arijit *et al.* 1982)[3], poniendo énfasis en estimadores sesgados e insesgados.

Estos agrupamientos, además de representar una estructura específica de datos, constituyen unidades de análisis que tienen interés por sí mismas.

Esto es, debido a la importancia que tiene considerar las variables medidas en los distintos niveles de la jerarquía es que permite entender la relevancia de los participantes en el estudio. De ahí, que el término multinivel (Goldstein, 2002)[13] se utilice tanto para designar los niveles en que se pueden agrupar los datos y para caracterizar a las formas o técnicas utilizadas para modelar relaciones que se dan dentro y a través de los diferentes niveles.

Dicha estructura de jerarquización puede ser analizada utilizando modelos, ya sea desde el punto de vista clásico como Bayesiano, es este último en el que se trabajó para realizar la estimaciones en poblaciones infinitas, con datos t-Student pudiéndose extender a otras distribuciones.

Además esta teoría tiene como propósito flexibilizar el modelamiento bajo supuestos de independencia y/o idéntica distribución. No obstante el apoyo de los avances tecnológicos, la rapidez de procesamiento y generación de algoritmos estadísticos, ha permitido que variadas investigaciones se centren y se sigan desarrollando en el análisis y modelamiento de estos, desde un enfoque paramétrico (Ortega-Basulto, 2009)[17], (Bush, 1996)[7], como no paramétrico (Casanova, 2007)[11], (Orellana, 2002)[18], (Ranganathan, 2004)[2].

Diversos trabajos se han realizado bajo el contexto de familias paramétricas que contengan una amplia gama de distribuciones, las cuales posean algunas características como por ejemplo, simetría o asimetría, colas más o menos pesadas (Thomson, 2005)[20], de tal manera de poder extender el modelamiento y tener algún grado de flexibilidad. Es así que este trabajo adoptará alguna de ellas y presentará algunos modelos bajo una estructura de correlación de manera que datos con las características mencionadas anteriormente puedan ser modelados.

En diferentes investigaciones desarrolladas en poblaciones finitas, es usual que se focalicen en estimaciones de medias, percentiles, varianzas, entre otras; todas estas desde el punto de vista de modelos lineales, (Bolfarine *et al.*, 1992)[6], (Bellhouse, 1987)[5] las cuales presentan resultados de estima-

dores sesgados e insesgados; dichas investigaciones contemplan resultados equivalentes entre muestreo aleatorio simple y la noción de variables aleatorias permutables, lo cual nos permite realizar inferencia a partir de un buen diseño de muestreo (Ericson, 1969)[12] y extender dichos resultados a poblaciones infinitas.

Prácticamente hasta la década de los 80, la construcción de modelos conjugados y el análisis asintótico, constituían las dos únicas formas de análisis Bayesiano. Así, junto con los desarrollos conceptuales en la formulación de modelos y el aumento de la tecnología computacional, se ha dado un nuevo impulso a los modelos bayesianos, cuya versatilidad para modelar situaciones más complejas no habían sido explotadas debido a resolución de procesos analíticos imposibles de resolver.

En los tiempos actuales se han implementado métodos de integración y de simulación; luego el marco Bayesiano parece ser el método a seguir para el análisis de modelos de este tipo, además el auge de las técnicas y/o métodos de simulación llamados en su generalización métodos MCMC, es decir, Métodos Montecarlo con Cadenas de Markov, se puede lograr en forma relativamente simple obtener muestras de una distribución objetivo que junto con los trabajos de Metropolis, permite dar fluidez y eficacia a la obtención de muestras de distribuciones a posteriori complejas, sobre todo cuando se trabaja con alta dimensionalidad.

Enfoques Estadísticos

El Objetivo del análisis estadístico es la inferencia, que intuitivamente se puede definir como el proceso de aproximar o estimar qué $P \in \boldsymbol{P}$ genera los datos o bien, verificar o refutar alguna hipótesis acerca de la verdadera medida P.

Dependiendo de la naturaleza de la familia de medidas de probabilidad $\boldsymbol{P}$, es posible distinguir tres enfoques en la inferencia: el enfoque paramétrico, el no-paramétrico y el semi-paramétrico.

La inferencia paramétrica considera que cada medida de probabilidad de la familia, está indexada por un parámetro real θ , escalar o vectorial. En tal caso escribimos: $\boldsymbol{P} = \{P_\theta : \theta \in \vartheta\}$, donde ϑ es el llamado espacio de parámetros, el cual es finito-dimensional. Es usual suponer, en este caso, que a cada medida de probabilidad se le puede asociar una función de densidad (o cuantía), indexada por θ . De esta forma surgen los diferentes modelos estadísticos paramétricos tales como los modelos Bino-

mial, Poisson, Exponencial, Normal, t-Student, Slash, etc.

La inferencia no-paramétrica se refiere al caso cuando no se supone forma alguna para las medidas de probabilidad de la familia P, en tal caso puede considerarse como conjunto de índices al conjunto de todas las funciones de distribución F de interés, el cual es usual denotarlo por $\varrho = \{P_F : F \in F\}$, así el espacio de parámetros es infinito-dimensional.

Finalmente, se ha adoptado en llamar modelos estadísticos semi paramétricos, a aquellos modelos que toman en consideración, tanto parámetros finito dimensionales, como infinitos dimensionales, considerando habitualmente una estructura jerárquica en su definición.

Modelo Bayesiano

El modelo que se desarrolló en este trabajo y sus posibles extensiones son modelos multivariantes con una estructura jerárquica Bayesiana, es así, que se mencionarán algunas características y conceptos que son utilizados en ellos.

Un modelo estadístico es un trío formado por: $(Y; A; P_\theta)$ donde Y, es el espacio de posibles observaciones, A es la σ -álgebra asociada a Y y P_θ es una familia de posibles medidas de probabilidad indexadas por un parámetro θ y definidas sobre $(Y; A) = \Omega$.

Definición.

El modelo Bayesiano en su forma más sencilla consiste de un parámetro o vector de parámetros desconocido θ aleatorio, con $\theta \in \vartheta$, donde ϑ es el espacio paramétrico de valores para θ , el cual también tiene asociado una σ -álgebra y una distribución a priori $\prod$ de ϑ , perteneciente a alguna familia de distribuciones posibles, con una medida de probabilidad asociada, dada por el conocimiento o experiencia del experto.

Una forma simple y útil de visualizar el modelo Bayesiano es en dos etapas o jerarquías, la primera corresponde al modelo observacional o verosimilitud y la segunda a la distribución a priori, es decir:

$$y|\theta \sim p(y|\theta).$$
$$\theta \sim \pi(\theta).$$

Los modelos jerárquicos son preferidos por diferentes razones, entre ellas están:

1. En análisis de meta poblaciones, es decir en estudios de aquellas poblacio-

nes que a su vez se subdividen en subpoblaciones.

2. En el ámbito no informativo, los modelos jerárquicos le dan robustez a las estimaciones (Berger, 1985)[4] y en muchos casos se trabaja con distribuciones no informativas relacionadas con distribuciones conjugadas.

3. Un tercer caso, la formulación jerarquizada puede flexibilizar el procedimiento computacional.

Métodos Computacionales

La inferencia en los modelos Bayesianos radica sustancialmente en la información que la distribución a posteriori nos entregue, entre ellas podemos mencionar: medidas de tendencia central, medidas de variabilidad, intervalos de credibilidad, distribuciones marginales, factores de Bayes, etc; dichos resultados involucran cálculo de Integrales que tienen la misma dimensión de los parámetros a estimar, con lo cual la dificultad para obtener los resultados se vuelve compleja; es así, que se han desarrollado mecanismos para simplificar estos procesos. Entre ellos destacaremos el algoritmo de Gibbs Sampling y el de Metropolis Hasting (casos particulares de MCMC).

El método de Gibbs debe su popularidad, al hecho que en muchos modelos estadísticos la distribución condicional a posteriori completa

$f(\theta_j \mid y, \theta_k, k \neq j)$ es posible de simular. Ocurre sin embargo, casos en donde esto no es posible y por ello se hace necesario contar con otros métodos MCMC alternativos (posiblemente el más genérico de estos esquemas es el de Metropolis).

El muestreador de Gibbs y el esquema de Metropolis-Hasting son por construcción invariantes con respecto a la distribución a posteriori buscada. Lo que uno debe verificar entonces son la aperiodicidad e irreducibilidad de la cadena, siendo esta última la más crítica pues en ocasiones es posible encontrar un subconjunto de estados, tales que, la cadena simulada entre en ellos sea improbable salir y el algoritmo por tanto se entrampe en ese punto sin llegar a converger. En la práctica, más importante que establecer convergencias teóricas, es reconocer la convergencia práctica; es decir juzgar cuantas transiciones M debe de ser suficientes como para obtener promedios ergódicos que estén cerca de la media deseada. Una vez finalizada esta etapa se comienza el proceso de Inferencia para el o los parámetros o modelos en cuestión.

Comparación de Modelos

Si se desea comparar varios modelos para determinar cuál de ellos posee mayor precisión en sus estimaciones o en su capacidad de predicción, existen variados métodos, dentro de ellos podemos mencionar: Criterio de Información Bayesiana (BIC), Criterio de Información de Akaike (AIC), Criterio de Información de Devianza (DIC),y Criterio de Mínima Pérdida Predictiva a Posteriori (D), entre otros, en particular se usó el BIC, el DIC y pd para realizar las comparaciones de modelos y sus grados de complejidad.

Modelos Bayesianos Bajo Independencia

El muestreo en varias etapas es frecuentemente usado en análisis de poblaciones humanas, (Kish, 1965)[15], describe un muestreo en tres etapas de extracción, en el cual una muestra de países es extraída en una primera etapa, una muestra de bloques (estratos) es extraída de la muestra de cada país como segunda etapa, y luego de cada bloque, se extrae una muestra como tercera etapa. Además, en cada etapa se utiliza un muestreo aleatorio simple, y se procede a realizar todas las estimaciones de las cantidades deseadas de acuerdo al interés del estudio.

En este trabajo se asume que el diseño del muestreo y la obtención de la muestra fue previo al análisis que sigue, de tal manera que se cuenta con k subpoblaciones de un total de K y de cada una de ellas fue extraída una muestra de tamaño n_i, completando un total de n observaciones muestrales de la superpoblación, es decir, $\sum_{i=1}^{k} n_i = n$.

Se utilizaron algunos supuestos iniciales basados en un trabajo realizado por Scott & Smith (1969)[19], cuyo objetivo es realizar estimaciones en poblaciones finitas. No obstante este trabajo se abordó desde el punto de vista de poblaciones infinitas, donde el muestreo se desarrolla en varias sub-etapas, con una estructura de correlación dentro de las subpoblaciones y no correlación fuera de ellas.

La variable de interés se denotó por y_i con n_i elementos muestrales, donde, $y_{i1},\ldots,y_{in_i}$, denota una muestra de la i-ésima subpoblación.

Para especificar el modelo para la población de elementos, el diseño estará dado bajo los siguientes supuestos:

S1: Los elementos del i-ésimo cluster, están no correlacionados y tienen distribución Fi, con media μ_i y varianza $\sum_i$.

S2: Las medias entre los cluster están no correlacionadas y tienen distribución G con media ν y varianza δ^2.

Estos supuestos son equivalentes a especificar una superpoblación, desde la cual, una población finita ha sido extraída y puede ser expresada en términos del concepto de permutabilidad de de Finetti's (Ericson, 1968)[12].

En el modelo planteado se presenta la permutabilidad entre los elementos de cada subpoblación y entre las medias de los clusters.

Considerando lo anteriormente expuesto, en este trabajo la distribución F está dada por una mezcla de distribuciones normales, para dar mayor flexibilidad al modelamiento. G es una distribución normal y el parámetro de localización de esta última, en algunos casos será una constante conocida y en otros se asumirá como variable aleatoria con una distribución a priori normal, con media cero y varianza grande (mayor o igual a 1000).

No obstante, el interés principal de este trabajo fue estimar las medias y varianzas de cada subpoblación.

Modelo normal heterocedástico con variable auxiliar

Uno de los fines principales de este trabajo es la de poder realizar inferencia en subpoblaciones con observaciones modeladas con distintas distribuciones. De esta manera, se introduce una variable auxiliar u_i la que nos permite trabajar con mezcla de normales, así, la forma de la variable aleatoria que modela los datos, dados los parámetros de interés que son la media y la varianza, podrá ser representada estocásticamente como:

$$\mathbf{y}_i \overset{d}{=} u_i^{-\frac{1}{2}} \boldsymbol{\Sigma}_i^{\frac{1}{2}} \cdot \mathbf{z}_i + \boldsymbol{\mu}_i,$$

tal que u_i es independientes de z_i, u_i distribuye G con G(0) = 0.

Modelo t-Student heterocedástico

Al flexibilizar las varianzas asumiéndolas distintas para cada subpoblación, modelando sus distribuciones a priori como IGamma con parámetros a_i, b_i, respectivamente y modelando las u_i

con una Gamma. Se obtiene el siguiente modelo jerárquico:

$$\begin{aligned} y_{ij} \mid \mu_i, \sigma_i^2, u_i &\overset{c.iid}{\sim} N(\mu_i, u_i^{-1}\sigma_i^2) \\ \mu_i \mid \nu, \delta^2 &\overset{iid}{\sim} N(\nu, \delta^2) \\ \sigma_i^2 \mid a_i, b_i &\sim IG(a_i, b_i) \\ u_i \mid \varphi &\overset{iid}{\sim} G(\frac{\varphi}{2}, \frac{\varphi}{2}). \end{aligned}$$

Donde $\delta^2, a_i, b_i, \varphi$ son constantes reales conocidas y positivas, $\varphi > 2$, de $i = 1,...,k$.

Luego las distribuciones condicionales para el algoritmo Gibas Sampling están dadas por:

Algoritmo 1

$$\begin{aligned} \mu_i \mid \mathbf{y}_i, \sigma_i^2, u_i, \nu, \delta^2 &\sim N(A_i, B_i) \\ \sigma_i^2 \mid \mathbf{y}_i, \mu_i, a_i, b_i, u_i &\sim IG\left(a_i + \frac{n_i}{2}, S^{(i)}\right) \\ u_i \mid \mathbf{y}_i, \mu_i, \sigma_i^2, \varphi &\sim G\left(\frac{n_i + \varphi}{2}, C_i\right). \end{aligned}$$

Donde A_i, B_i, C_i, están especificadas en la sección anterior reemplazando σ^2 por σ_i^2 y

$$S^{(i)} = \frac{u_i \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2}{2} + b_i.$$

Modelo Bayesiano Correlacionado

En esta etapa se trabajó el modelo propuesto anteriormente estableciendo la existencia de correlación al interior de cada subpoblación y no correlación entre ellas.

Para especificar el modelo para la población de elementos, el diseño estará dado bajo los siguientes supuestos:

S1: Los elementos del i-ésimo cluster, están correlacionados, con distribución Fi, de media μ_i y varianza $\sum_i$.

S2: Las medias $\mu_1,...,\mu_k$, están no correlacionadas, con distribución G, de media ν y varianza δ^2.

Tenemos que $E(y_i) = \mu_i$ y

$$\begin{aligned} Cov(y_{ij}, y_{kl}) &= c_i + \sigma_i^2 && i = k, j = l, \\ &= c_i, && i = k, j \neq l \\ &= 0, && i \neq k \end{aligned}$$

Donde c_i actúa como una medida de dispersión, la que asumirá un valor real positivo y por lo tanto nuestra matriz de varianzas-covarianzas queda de la forma:

$$\Sigma_n = \begin{bmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & \\ & & \ddots & \\ & & & \Sigma_k \end{bmatrix} \quad , \quad \Sigma_i = \sigma_i^2 \mathbf{I}_{n_i} + c_i \mathbf{J}_i \quad , \mathbf{J}_i = 1_{n_i} 1_{n_i}^t .$$

Dicha estructura, presenta algunas dificultades para obtener las distribuciones a posteriori y desde luego realizar las estimaciones. Es así, que además de la variable auxiliar u_i de la sección anterior, se introduce una variable aleatoria w_i, para incluir el factor de covarianza en el modelamiento y poder aplicar el esquema de Gibas Sampling.

La variable w_i actuará como una componente aditiva a la media del modelo de los datos y en el modelamiento de esta, se incluirá la covarianza c_i como una medida de dispersión, con el objetivo de poder realizar las estimaciones correspondientes bajo esta estructura.

Se asumió que $w_{ij} \coprod \mu_i$, $\forall\, i = 1,...,k,\ y\ \forall\, j = 1,...,n_i$; además la variable aleatoria w_{ij} distribuye normal con media 0 y varianza $u_i^{-1} c_i$, de $i = 1,...,k,\ j = 1,...,n_i$. Para todos los efectos, w_i denotará el vector de componente aditiva a los datos de dimensión n_i.

Así, la media de la distribución marginal para yi para el modelo t-Student bajo una estructura de correlación está dada por:

$$E(\mathbf{y}_i) = E(E(\mathbf{y}_i|\boldsymbol{\mu}_i, \mathbf{w}_i, u_i)) = E(\boldsymbol{\mu}_i + \mathbf{w}_i) = \boldsymbol{\mu}_i.$$

cuya varianza es:

$$\begin{aligned} V(\mathbf{y}_i) &= V(E(\mathbf{y}_i|\boldsymbol{\mu}_i, \mathbf{w}_i, u_i)) + E(V(\mathbf{y}_i|\boldsymbol{\mu}_i, \mathbf{w}_i, u_i)) \\ &= V(\boldsymbol{\mu}_i + \mathbf{w}_i) + E(u_i^{-1}\sigma_i^2 \mathbf{I}_{n_i}) \\ &= \frac{\varphi/2}{\varphi/2-1}\Sigma_i. \end{aligned}$$

Condiciones para la covarianza son las siguientes:

1. Se asume que entre dos variables hay correlación positiva ρ, si existe una constante positiva ξ, tal que: $\xi < \rho < 1$. Un criterio estándar, para que exista correlación positiva significativa entre dos variables, es que esta se

encuentre entre 0,5 y 1 (a menos que se ocupe un criterio más amplio), luego entre las varianzas y la covarianza se debe cumplir que, $0 < \sigma_i^{\,2} < \frac{1-\xi}{\xi} \cdot c_i$ para subpoblaciones heterocedásticas con covarianzas distintas o que él $\max\{\sigma_i^{\,2}\} < \frac{1-\xi}{\xi} \cdot c_i$; para covarianzas iguales entre las subpoblaciones, $\forall\, i = 1,...,k.$

2. En el caso de querer ampliar el modelamiento respecto de la covarianza, es decir, en caso de desear analizar covarianza negativa, la restricción para estas es: $-\frac{\sigma_i^{\,2}}{2} < c_i < -\frac{\xi}{1+\xi} \cdot \sigma_i^{\,2}$, o

$$\max\{-\sigma_i^{\,2}/2\} < c < \min\{-\frac{\xi}{1+\xi} \cdot \sigma_i^{\,2}\}$$

, $\forall\, i = 1,...,k,$ con $0 < \xi < 1.$ De esta forma se propone como solución para las estimaciones, el mecanismo MCMC.

3. Si la varianza es muy grande, la covarianza podría flexibilizar su rango, admitiendo la posibilidad de modelarla a través de una distribución normal en el esquema Bayesiano.

Modelo t-Student heterocedástico

En este modelo se asume la aleatoriedad de la variable auxiliar u_i, a través de un modelo Gamma al igual que en el caso anterior, y la variable w_i especificada, de esta manera se obtiene una distribución t-Student para la marginal y_i la que se expresa de la siguiente manera:

$$\mathbf{y}_i \mid \mu_i, \Sigma_i, \varphi, c_i \overset{c.ind}{\sim} t(\mu_i \mathbf{1}_{ni}, \Sigma_i, \varphi).$$

Donde

$$\Sigma_i = c_i \mathbf{1}_{n_i} + \sigma_i^2 \mathbf{I}_{n_i}.$$

Luego el modelo jerarquizado bajo una estructura de correlación está dado por:

$$\begin{aligned}
y_{ij} \mid u_i, \mu_i, w_{ij}, \sigma_i^2, c_i &\overset{c.iid}{\sim} N(\mu_i + w_{ij}, u_i^{-1}\sigma_i^2) \\
w_{ij} \mid u_i, c_i &\overset{c.ind}{\sim} N(0, u_i^{-1} c_i) \\
\mu_i \mid \nu, \delta^2 &\overset{iid}{\sim} N(\nu, \delta^2) \\
u_i \mid \varphi &\overset{iid}{\sim} G(\varphi/2, \varphi/2) \\
c_i \mid e_i, f_i &\sim IG(e_i, f_i) \\
\sigma_i^2 \mid a_i, b_i &\sim IG(a_i, b_i).
\end{aligned}$$

Con $a_i, b_i, e_i, f_i, \delta, \varphi \in IR^+, \nu \in IR$ constantes conocidas.

Algoritmo 2

$$\begin{aligned}
\mu_i \,|\mathbf{y}_i,\, \sigma_i^2,\, \nu,\, \delta^2,\, \mathbf{w}_i,\, u_i &\sim N\,(\,A_i\,,\,B_i) \\
w_{ij} \mid y_{ij},\, \mu_i,\, u_i,\, c_i,\, \sigma_i^2 &\sim N\,(\,\kappa_{ij}\,,\,\tau_{ij}) \\
\sigma_i^2 \,|\mathbf{y}_i,\, a_i,\, b_i,\, u_i,\, \mu_i,\, \mathbf{w}_i &\sim IG\left(\,\frac{n_i}{2}+a_i\,,\,S^{(i)}\right) \\
u_i \,|\mathbf{y}_i,\, \mu_i,\, \mathbf{w}_i,\, \varphi,\, c_i &\sim G\left(\,\frac{\varphi}{2}+n_i\,,\,B1_i\right) \\
c_i \,|u_i,\, \mathbf{w}_i &\sim IG\left(\,\frac{n_i}{2}+e_i\,,\,H_i\right).
\end{aligned}$$

Por lo tanto las distribuciones Condicionales completas obtenidas para el algoritmo Gibbs Sampling son:

Con,

$$A_i = \frac{n_i \cdot \delta^2 \cdot u_i \cdot (\overline{\mathbf{y}_i} - \overline{\mathbf{w}_i}) + \nu \cdot \sigma_i^2}{n_i \cdot u_i \cdot \delta^2 + \sigma_i^2}, \quad B_i = \frac{\sigma_i^2 \cdot \delta^2}{u_i \cdot n_i \cdot \delta^2 + \sigma_i^2}$$

$$\kappa_{ij} = \frac{(y_{ij} - \mu_i) \cdot c_i}{c_i + \sigma_i^2}, \quad \tau_{ij} = \frac{\sigma_i^2 \cdot c_i}{u_i \cdot (c_i + \sigma_i^2)}$$

$$S^{(i)} = b_i + \frac{1}{2} \cdot \left(u_i \, \Sigma_{j=1}^{n_i} \left(y_{ij} - (\mu_i + w_{ij})\right)^2\right),$$

$$B1_i = \frac{\frac{||\mathbf{y}_i - (\mu_i \, \mathbf{1}_{n_i} + \mathbf{w}_i)||^2}{\sigma_i^2} + \frac{||\mathbf{w}_i||^2}{c_i} + \varphi}{2}, \quad H_i = \frac{u_i \, ||\mathbf{w}_i||^2 + 2 f_i}{2}.$$

Simulaciones

El objetivo en esta etapa, es apoyarnos de procesos de simulación para ejecutar nuestros procedimientos y así poder inspeccionar las condiciones que debe poseer el sistema para realizar estimaciones apropiadas.

Las simulaciones son comunes hacerlas bajo varios escenarios, las que nos

permiten analizar comportamientos de los elementos involucrados, medir la efectividad de los procesos respecto de condiciones dadas y evaluar su factibilidad.

Estas permiten medir el impacto de cambios en el proceso, mejora el conocimiento de este y enriquece la toma de decisiones, sobre todo cuando se les quiere aplicar a situaciones reales.

Dentro de los elementos claves que ayudan en el desarrollo de un modelo de simulación, se encuentran:

1. Tamaño de corridas adecuadas.

2. Especificaciones claras de las variables involucradas.

3. Relaciones adecuadas entre los elementos del sistema.

4. Análisis constante del comportamiento de los resultados obtenidos.

5. Realizar los ajustes necesarios dentro de las especificaciones establecidas para comprobar la mejora de resultados.

6. Reportar debilidades y fortalezas del sistema.

En la ejecución de las simulaciones se consideraron algunos aspectos, dentro de los cuales se destacan:

1. Se realizaron variadas simulaciones, con condiciones iniciales arbitrarias con dos subpoblaciones.

2. Los valores de los hiperparámetros de las prioris fueron similares para establecer posibles comparaciones.

3. Las simulaciones se realizaron para cada uno de los modelos con tamaños muestrales distintos.

4. En general el número de iteraciones fue alto para garantizar convergencia de las cadenas.

5. Se eliminó un porcentaje inicial de la cadena, asumiendo un máximo dentro del análisis como criterio estándar (en algunos casos pudo ser menos), para disminuir el efecto de condiciones iniciales.

6. Una vez eliminado una cantidad inicial, del porcentaje restante de la cadena se realizó un muestreo sistemático, para obtener la convergencia de estas, utilizando los test de Raftery and Lewis, Heidleberger and Welch y Geweke. No obstante los resultados que se presentan en las tablas, están dados en general con muestreo cada 10, al utilizar el test de Raftery and Welch en algunos casos falla la convergencia de las varianzas.

Para los modelos independientes los valores de hiperparámetros que se optaron para la simulación, fueron:

$$\varphi = 10,\ a_1 = a_2 = 5,\quad b_1 = 7, b_2 = 6,\ \nu = 10,\quad \delta^2 = 100.$$

Con estos valores, la priori de las medias, quedó centrada en 10 con varianza 100, las prioris para las varianzas de las poblaciones 1 y 2, quedaron centradas en 1.75 y 1.5 con varianzas de 1.02 y 0.75 respectivamente.

Los valores de parámetros utilizados, para realizar las comparaciones versus las estimaciones fueron:

$$\mu_1 = -99{,}4423,\ \mu_2 = 60{,}9660,\ \sigma_1^2 = 1,\ \sigma_2^2 = 1.2.$$

En el modelo bajo Independencia se realizaron 30 veces 50 simulaciones para medir la calibración de estos (con 10000 iteraciones) con distintos tamaños muestrales, cuyo reporte final, se obtuvo con, ni = 30; 60; 100 observaciones, que se indica en el cuadro 1. Además bajo pérdida cuadrática, las estimaciones para los parámetros, junto con el error, están dadas por la media y la varianza de la distribución a posteriori (estimada), que se obtienen de las cadenas, junto con el resultado de su intervalo de credibilidad.

Las estimaciones que se presentan a continuación se realizaron con 100000 a 150000 iteraciones, de ellas se eliminaron el 20 por ciento inicial y se realizó un muestreo sistemático de las cadenas usando MCMC para realizar las estimaciones y/o analizar la convergencia de estas. Los largos que se emplearon fueron cada 10, 20 o 25 de ellas, valores extremos para que cada una de ellas presentara convergencia.

Obtenida la muestra de la iteración testeada se realizaron las estimaciones.

Estimaciones

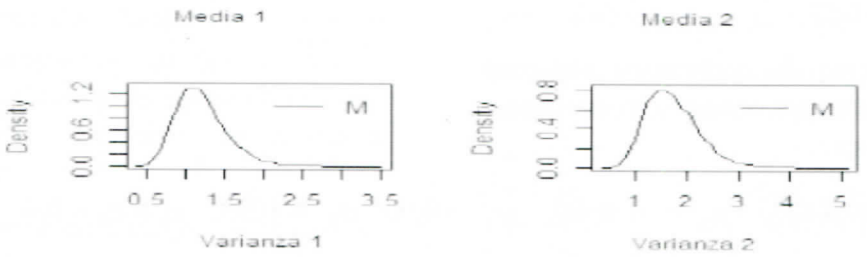
CUADRO 1: AJUSTE DE MODELO T-STUDENT IND.

Criterio	N° de Obs.	t-Student
DIC	30	186,3018
	60	400,3599
	100	642,2035
Pd	30	1,241
	60	0,7341
	100	0,4998
BIC	30	207,7842
	60	425,6301
	100	653,3115

CUADRO 2: ESTIMACIONES CON MODELO T-STUDENT

Parámetro		Nº de Obs.	Media	Desv. est	IC	
μ_1	-99,44233	30	-99,2633353	0,127438456	-99,5240903	-99,015602
		60	-99,373233	0,077265866	-99,5291837	-99,227343
		100	-99,475404	0,088414478	-99,6423434	-99,293704
μ_2	60,96605	30	60,8641464	0,166189681	60,5399535	61,200805
		60	61,061092	0,0989714	60,8658511	61,25798
		100	60,980657	0,092039633	60,7970204	61,161044
σ_1^2	1	30	0,7688306	0,217308352	0,4161809	1,202117
		60	1,176835	0,312293324	0,6327922	1,808733
		100	1,376123	0,39592596	0,6719182	1,877684
σ_2^2	1,2	30	1,4569299	0,470822026	0,6671733	2,435565
		60	1,590704	0,475130675	0,769715	2,557025
		100	1,424064	0,422127737	0,6673381	2,22171

FIGURA 1: D. POSTERIORI EN EL MODELO T-STUDENT



Estimaciones bajo un Modelo con observaciones correlacionadas

Dado que los procesos desarrollados en el capítulo anterior, se asumieron con covarianza positiva, y considerando las relaciones con las varianzas respectivas, se consideró un valor extremo inferior para ξ de 0.38. Es así, que los valores de los hiperparámetros que se utilizaron en esta etapa, en algunos casos son similares al caso Independiente, con algunas variaciones para mejorar las estimaciones de los mismos, con una estructura de correlación.

Cuadro 3: Ajuste de Modelo Correlacionado

DIC	30	170,093
	60	**317,0633**
	100	565,3868
Pd	30	0,6632
	60	0,5783
	100	0,5052
BIC	30	204,7153
	60	**351,8552**
	100	600,325

Los valores utilizados para los Hiperparámetros de la distribución a priori de las medias fue de ; por lo tanto la priori

$$\delta^2 = 100,\ \nu = 0,\quad e_1 = e_2 = 4,\quad f_1 = f_2 = 3,\quad a_1 = a_2 = b_1 = b_2 = 5,\quad \varphi = 4,$$

de las medias quedó centrada en 0 con varianza 100 y las prioris de las varianzas quedaron centradas en 1.25 con varianza 0.52.

Los valores de parámetros utilizados para realizar las comparaciones respectivas fueron:

$$\mu_1 = -99{,}4423,\quad \mu_2 = 60{,}9660,\quad \sigma_1^{\ 2} = 0{,}9,\quad \sigma_2^{\ 2} = 0{,}95,\quad c_1 = c_2 = 0{,}6.$$

Se realizaron 50 simulaciones (30 veces) con 10000 iteraciones para medir la calibración de este, cuyo reporte se observa en el cuadro 3.

CUADRO 4: ESTIMACIONES CON MODELO T-STUDENT

Parámetro		N° de Obs.	Media	Desv. est	IC	
μ_1	-99,4423273	30	-99,1823711	0,641901616	-100,3735173	-97,8819423
		60	-98,9803076	0,420805623	-99,7901049	-98,1523699
		100	-99,404674	0,259079766	-99,9504046	-98,9303442
μ_2	60,9660504	30	59,7709879	0,628226425	58,5638597	61,0310074
		60	60,8220986	0,388038498	60,024791	61,5408557
		100	60,5782231	0,340136698	59,9068715	61,2510049
σ_1^2	0,9	30	1,0972895	1,261788849	0,2386109	2,347862
		60	0,9729415	0,420100321	0,3888181	1,785174
		100	0,9160923	0,380649156	0,3348848	1,656621
σ_2^2	0,95	30	1,0598901	0,491689129	0,3648663	1,934941
		60	0,958282	0,40782983	0,35908	1,755087
		100	0,9689479	0,475496425	0,3270357	1,861962
c_1	0,7	30	0,682589	0,343030319	0,2244977	1,349511
		60	0,6793544	0,307296713	0,2287144	1,273935
		100	0,6403724	0,269277069	0,2177027	1,163145
c_2	0,6	30	0,7389347	0,344017732	0,2380541	1,391009
		60	0,6819921	0,305641718	0,240162	1,313654
		100	0,6752505	0,325223062	0,2173254	1,322658

Conclusiones generales

Una conclusión fundamental que se puede establecer, es que el uso de técnicas Bayesianas permiten resolver con cierta comodidad un problema complejo.

En el presente trabajo, se realizó un análisis bayesiano para poder obtener estimaciones de parámetros desconocidos, en una primera etapa bajo el supuesto básico de no correlación y como segunda etapa bajo el supuesto de correlación entre observaciones de subpoblaciones independientes entre ellas, no obstante, se realizaron ambos casos para establecer comparaciones.

En ambos modelos expuestos, se utilizaron distribuciones a priori usuales (familia conjugada) para los parámetros de interés, como también para las variables utilizadas para realizar el proceso, dentro de las cuales están las distribuciones normales para modelar medias y la distribución gamma o Igamma para modelar varianzas.

El modelos asociado a los datos fue t-Student, no obstante es posible trabajar de manera simple con modelo Normal, y slash, dando un grado de flexibilidad al proceso. Se obtuvieron las funciones de verosimilitud y las ecuaciones para aplicar Gibas Sampling, esto último dada la dificultad analítica en obtener las distribuciones a posteriori de los parámetros. De esta manera se pudo desarrollar la implementación de los algoritmos establecidos y así mediante el muestreo de las densidades a posteriori generadas por las cadenas vía simulación, obtener las estimaciones de los estimadores de máxima verosimilitud (estimadores de Bayes).

Las estimaciones obtenidas mediante los supuestos de independencia y bajo correlación son igualmente satisfactorias, claro está que cuando se aumenta el tamaño muestral el(o los) modelos bajo independencia, tienden a ser más óptimas.

Los resultados de calibración de los modelos independientes resultaron adecuados, no obstante en los modelos bajo correlación dicha calibración se presenta menos eficiente. Cabe hacer notar que el objetivo es detectar correlación y/o estimar correlaciones entre observaciones, por lo cual el mecanismo se vuelve más complejo en su implementación y por tanto las simulaciones deben ser más trabajosas en el proceso.

Dentro de un proceso de Inferencia usual, aplicada a datos reales, uno podría estar interesado en determinar la independencia de las observaciones (no obstante depende de la investigación en sí), claro está, que para hacer análisis de correlación entre observaciones en el método clásico, habrá que tener muchas réplicas de ellas, lo cual no es usual por diferentes motivos. Así, el enfoque Bayesiano nos permite determinar o analizar la existencia o no de dicha correlación con la misma muestra de observaciones.

Bibliografía

– **Aggarwal**, J.K. Sample surveys: inference and analysis, Volumen 2.

– **Ananth**, Ranganathan (20th September 2004). The Dirichlet Process Mixture (DPM) Model.

– **Arijit**, Chaudhuri and **Raghunth**, Arnab (1982). On Unbiased Variance-Estimation with Various Multi-stage Sampling Strategies.The Indian Journal of Statistics, Vol. 44, Series B, Pt.1. pp.92-101. Indian Statistical Institute.

– **Berger**, James O.(1985). Statistical Decision Theory and Bayesian Analysis. Segunda Edición, Springer-Verlag.

– **Bellhouse**, D. R. (1987). Model- Based Estimation in Finite Population Sampling. The American Statistician, Vol. 41, No4(Nov., 1987). pp.260-262.

– **Bolfarine**, H.; **Zacks**, S. (1992). Prediction Theory For Finite Populations- A Case Study Approach.Springer-verlag.

– **Bush** C. and **MacEachern** S.(1996). A semiparametric Bayesian model for randomised block designs. Biometrika 83(2):275-285.

– **Cannon**, M.J.; **Warner**, L.; **Taddei**, J.A.; **Kleinbaum**, D.G. (2001). What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in Brazil. Statistics in Medicine 20(9-10):1461-7.

– **Cnaan**, A.; **Laird**, N.M.; **Slasor**, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Statistics in Medicine 16(20):2349-2380.

– **Caseros**, Antonio (2010). Factores Modulares de la Percepción de la Calidad Docente. Journal of Educational Research, Assessment and Evaluation, Vol.71.No2.pp 90-94.

– **Casanova Laudien**, María Paz (2005). Análisis Bayesiano Semi paramétrico del problema de calibración en Modelos de Regresión Elípticos. Tesis de Doctorado en Estadística. Pontificia Universidad Católica de Chile.

– **Estimation** and inference using mixtures. Journal of American Statistical Association 85, 378-409.

– **Ericson**, W.A. (1969.a). Subjective Bayesian Models in Sampling Finite Populations. Journal of the Royal Statistical Society, Series B, (Methodological), Vol.31, No2. pp 195-233.

– **Goldstein**, H. (2002). Multilevel statistical models (3rd ed.). London: Arnold.

– **Ita Kreft** and **Jaude Leeuw** (jun 18, 1998). Introduction to multilevel Modeling.

– **Kish**, L. (1965). Survey Sampling. New York: John Wiley and Sons.

– **Legry** A.C.; **Donnan**, P.T.; **MacDonald**, T.M.; **Murphy**, M.B. (2000). The influence of physial activity on the variability of ambulatory blood pressure. American Journal of the Hypertension. 13 (10):1067-73.

– **Ortega Irizo**, Francisco Javier y **Basulto Santos**, Jesús. (2009). Estimación Bayesiana en Modelos de Producción con Frontera determinista.

– **Orellana Zapata**, Yasna (2007). Análisis Bayesiano No paramétrico utilizando Procesos Skew Dirichlet. Tesis de Doctorado en Estadística. Pontificia Universidad Católica de Chile.

– **Scott**, A. J. and **Smith**, T. M. F. (sep., 1969), Estimation in Multi-Stage Surveys. Journal of the American Statistical Association, Vol.64, No.327 pp. 830-840.

– **Thomson**, R. (1986). Estimation of realized heritability in a selected population using mixed model methods. Génét. Sél. Evol. 18(4): 475-484.