

**UNIVERSIDAD MAYOR DE SAN ANDRES  
FACULTAD DE CIENCIAS PURAS Y NATURALES  
CARRERA DE ESTADÍSTICA**



**TESIS DE GRADO**

**“MODELACIÓN POISSON CON ENFOQUE BAYESIANO  
Y ENFOQUE CLÁSICO, PARA EXPLICAR EL  
DESCENSO DE LA MORTALIDAD EN BOLIVIA  
DURANTE LOS PRIMEROS AÑOS DE VIDA”**

POSTULANTE: PATRICIA LOZA CRUZ

TUTOR: M. Sc. RAMIRO COA CLEMENTE

LA PAZ - BOLIVIA  
2013

## **DEDICATORIA**

*A mis adorados padres Arturo y Emma,  
por el ejemplo de vida. A mis  
Herman@s: Sandra Susy, Marisol,  
Angel Arturo, Percy, Fernando, Manuel  
Emilio, Ana Luisa, Mauricio y Limber  
por su impulso constante. A Ramiro, por  
su apoyo y paciencia.*

## AGRADECIMIENTOS

Deseo expresar mi más sincera gratitud a mi tutor por el apoyo tan profesional brindado en el trascurso del desarrollo de la presente tesis, por darme la posibilidad de compartir ideas e inquietudes sobre las problemáticas existentes en el área de Demografía, Programación y Estadística Bayesiana, ya que me permitió comprender aún más las múltiples complejidades del presente trabajo.

A todo el entorno de la Facultad de Ciencias Puras y Naturales, administrativos, profesores y compañeros.

Finalmente, cabe un agradecimiento muy sincero y particular a la señora bibliotecaria de la carrera de Estadística Lidia Daza por su colaboración y simpatía durante toda mi estadía universitaria, en cuyo trabajo trasluce su pasión por la carrera de Estadística, su entorno y cultura.

## RESÚMEN

Un modelo de regresión Poisson es adecuado para abordar el análisis del descenso del nivel de la mortalidad en los primeros años de vida así como también los factores determinantes. Afrontarlo desde el enfoque Clásico o enfoque Bayesiano conlleva a una serie de complejidades, desde el planteamiento del modelo, hasta experimentar una sobredispersión aparente o real. La implementación de técnicas de estimación bayesiana en dichos métodos ha alcanzado especial relevancia recientemente debido a los avances en las técnicas de computación bayesiana, que permiten abordar de una forma relativamente cómoda los modelos.

Determinar la tendencia, la magnitud y el ritmo del descenso de la mortalidad neonatal, post-neonatal y post-infantil es de gran importancia para los planificadores del sector de la salud del país, pues la mortalidad infantil está estrechamente vinculada a la pobreza; debido a ello, los avances en la supervivencia de bebés y niños han sido más lentos en la población de los países pobres como Bolivia. La mejora de los servicios públicos de salud es un elemento clave, en particular el acceso a agua potable y a un mejor saneamiento. La instrucción de las madres puede salvar la vida de muchos niños. Si bien el aumento de los ingresos puede servir de algo, no será mucho lo que se consiga a menos que dichos servicios se presten a quienes más los necesitan.

El estudio tiene alcance a nivel nacional para infantes desde los 0 meses de edad hasta los 59 meses enmarcados desde el año 1993 hasta el año 2007, con base en la Encuesta Nacional de Demografía y Salud, principal fuente de

información del País con relación al sector salud. En el módulo de “historia de nacimientos” de la ENDSA 2008 se obtiene información acerca de la fecha de nacimiento, edad actual, y edad al morir si es el caso, entre otros datos, para cada uno de los hijos nacidos vivos de las mujeres en edad fértil entrevistadas, información que se usó en la aplicación de los enfoques propuestos.

## ÍNDICE

	Página
<b>INTRODUCCIÓN</b>	13
<b>1. ANTECEDENTES</b>	14
<b>2. PLANTEAMIENTO DEL PROBLEMA</b>	15
<b>3. HIPÓTESIS</b>	16
<b>4. OBJETIVOS</b>	16
4.1 OBJETIVO GENERAL	16
4.2 OBJETIVOS ESPECÍFICOS	16
<b>5. JUSTIFICACIÓN</b>	17
<b>6 ALCANCES Y LIMITACIONES</b>	21
<b>ENFOQUE CLÁSICO</b>	
<b>7. LA FAMILIA EXPONENCIAL</b>	22
7.1 DEFINICIÓN	22
7.2 MOMENTOS EN LA FAMILIA EXPONENCIAL	24
7.2.1 Condiciones de Regularidad	24
7.2.2 Momentos	25
7.2.3 Sobredispersión	30

<b>8. MODELOS LINEALES GENERALIZADOS</b>	31
8.1 COMPONENTES DE UN MODELO LINEAL GENERALIZADO (MLG)	31
8.2 ECUACIONES DE VEROSIMILITUD PARA EL MLG	33
8.3 EVALUACIÓN DE AJUSTE EN UN MLG	36
8.3.1 La función desvío	36
8.3.2 Análisis de Residuos	41
8.4 ESTIMACIÓN DE PARÁMETROS EN UN MLG	44
8.4.1 Método de Newton - Raphson	44
8.4.2 Método de Fisher Scoring	49
8.4.3 MCIR	49
8.4.4 Algoritmo de estimación	51

## ENFOQUE BAYESIANO

<b>9. PARADIGMA BAYESIANO</b>	56
9.1 FUNDAMENTOS DEL ENFOQUE	56
9.2 DESARROLLO DEL TEOREMA DE BAYES	57
9.2.1 Naturaleza secuencial del Teorema de Bayes	58
9.3 FUNCIÓN DE VEROSIMILITUD	60
9.4 DISTRIBUCIÓN PREVIA CONJUGADA	61
9.4.1 Definición	61
9.4.2 Distribución previa conjugada,	
Familia exponencial y suficiencia	62
9.5 DISTRIBUCIONES PREVIAS NO INFORMATIVAS	66
9.5.1 Principio de Invarianza de Jeffreys	67
9.6 DISTRIBUCIÓN PREDICTIVA	70
9.7 DISTRIBUCIÓN POSTERIOR	71
9.8 MÉTODOS MCMC	72

9.8.1 Algoritmo de Metrópolis - Hasting	74
9.8.2 Muestreador de Gibbs	79
9.9 MONITOREO DE CONVERGENCIA	81
9.9.1 Diagnóstico de Convergencia de Gelman y Rubin	82
9.9.1.1 Definición	82
9.10 COMPARACIÓN DE MODELOS	84
 <b>10. INFORMACIÓN</b>	 87
10.1 Estructura de la Información	87
 <b>11. RESULTADOS ENFOQUE CLÁSICO</b>	 90
11.1 Estimación del Modelo	90
11.2 Predicción y Análisis de Residuos	91
11.3 Efectos y Significancias	96
 <b>12. RESULTADOS ENFOQUE BAYESIANO</b>	 101
12.1 Estimación del Modelo	101
12.2 Diagnóstico de convergencia	102
12.3 Predicción	109
12.4 Efectos	111
 <b>13. CONCLUSIONES Y RECOMENDACIONES DE POLÍTICA</b>	 113
 <b>14. APÉNDICE</b>	 117

## BIBLIOGRAFÍA

## LISTA DE CUADROS

### **Cuadro 10.1**

Bolivia: Tasas de mortalidad observadas por tramo de edad,  
según cohorte de nacimiento, área de residencia y  
nivel de educación, ENDSA 2008\_\_\_\_\_ 88

### **Cuadro 11.1**

Estimación de los parámetros (en términos de razón de tasas de incidencia)  
para el modelo "edad\*cohorte + edad\*educación"\_\_\_\_\_ 90

### **Cuadro 11.2**

Efecto cohorte: Evolución de la mortalidad considerando  
cohortes extremas, por tipo de mortalidad\_\_\_\_\_ 97

### **Cuadro 11.3**

Efecto cohorte: evolución de la mortalidad  
considerando cohortes consecutivas, por tipo de mortalidad\_\_\_\_\_ 98

### **Cuadro 12.1**

Bayesiano: Estimación Bayesiana de los parámetros  
para el modelo elegido "EDA\*COH + EDA\*EDU + RES"\_\_\_\_\_ 101

### **Cuadro 12.2**

Bayesiano: Efecto cohorte: Evolución de la mortalidad  
considerando las cohortes extremas, por tipo de mortalidad\_\_\_\_\_ 112

### **Cuadro 12.3**

Bayesiano: Efecto cohorte: evolución de la mortalidad  
considerando cohortes consecutivas, por tipo de mortalidad\_\_\_\_\_ 112

## LISTA DE GRÁFICOS

### **Gráfico 5.1**

Bolivia: Evolución de las tasas de mortalidad Neonatal,  
post-neonatal y Post-infantil, combinando resultados de distintas ENDSAS\_\_\_\_\_19

### **Gráfico 5.2**

Bolivia: Evolución de las tasas de mortalidad Neonatal,  
post-neonatal y Post-infantil, según la ENDSA 2008\_\_\_\_\_20

### **Gráfico 10.1**

Bolivia: Tasas de mortalidad según cohorte de nacimiento, ENDSA 2008\_\_\_\_\_89

### **Gráfico 11.1**

Número de muertes observadas y muertes predichas  
con el modelo elegido "EDA\*COH + EDA\*EDU + RES"\_\_\_\_\_93

### **Gráfico 11.2**

Tasas de mortalidad observadas y tasas predichas  
con el modelo elegido "EDA\*COH + EDA\*EDU + RES"\_\_\_\_\_93

### **Gráfico 11.3**

Número de muertes observadas y muertes predichas  
con el modelo "EDA\*COH\*EDU + RES"\_\_\_\_\_94

### **Gráfico 11.4**

Tasas de mortalidad observadas y tasas predichas  
con el modelo "EDA\*COH\*EDU + RES"\_\_\_\_\_94

**Gráfico 11.5**

Número de muertes predichas y residuos basados en el criterio de devianza  
para el modelo elegido "EDA\*COH + EDA\*EDU + RES" \_\_\_\_\_ 96

**Gráfico 11.6**

Tendencia de la mortalidad en los primeros años de vida,  
una vez controlado el efecto de educación y residencia \_\_\_\_\_ 100

**Gráfico 12.1**

Diagnóstico de Convergencia para cada uno  
de los parámetros Modelo Eda\*Coh+Eda\*Edu+Res \_\_\_\_\_ 102

**Gráfico 12.2**

Historia de la convergencia de los parámetros. Evolución de la Simulación \_\_\_\_\_ 105

**Gráfico 12.3**

Densidad estimada de la Distribución Posterior \_\_\_\_\_ 108

**Gráfico 12.4**

Bayesiano: Número de muertes observadas y muertes predichas  
con el modelo elegido "EDA\*COH + EDA\*EDU + RES" \_\_\_\_\_ 110

**Gráfico 12.5**

Bayesiano: Tasas de mortalidad observadas y tasas predichas  
con el modelo elegido "EDA\*COH + EDA\*EDU + RES" \_\_\_\_\_ 111



## INTRODUCCIÓN

El estudio de la mortalidad en los primeros años de vida se basa en la observación de defunciones que ocurren en una población durante un tiempo determinado, conocer y comprender el ritmo de descenso de la mortalidad en los primeros años de vida es de relevante importancia para los decisores del sector Salud, pues ello permite evaluar la eficacia de acciones pasadas en el área, así como delinear nuevas políticas y programas. Empero, también es de vital importancia conocer los factores que determinaron tanto esos niveles de mortalidad como su tendencia<sup>1</sup>.

En el orden metodológico, un modelo de regresión Poisson es adecuado para abordar el análisis del descenso del nivel de la mortalidad en los primeros años de vida y los factores determinantes. Desarrollarlo desde el enfoque Clásico o enfoque Bayesiano conlleva el problema de sobredispersión típico en datos tipo Poisson, que puede superarse recurriendo o bien a un “mejor” tratamiento de la información en la etapa de modelación en el caso de sobredispersión aparente, o bien a un modelo de regresión Binomial Negativa en el caso de sobredispersión real. La modelación Poisson se desarrolló en torno a una teoría establecida clásica o frecuentista que hace uso exhaustivo de los datos observados; y se limita únicamente a la información que proporcionan éstos. Por otra parte, el manifiesto del paradigma bayesiano encauza a la estadística desde un panorama distinto, el de aprovechar tanto la información proporcionada por los datos como también la información extra muestral disponible, entendiendo por éste último toda aquella información relevante que ayude a disminuir la incertidumbre o ignorancia en torno a un fenómeno aleatorio de interés. La implementación de técnicas de estimación bayesiana en dichos métodos ha alcanzado especial relevancia recientemente debido a los

---

<sup>1</sup> Publicación de las Naciones Unidas, (1983). “Indirect Techniques for Demographic Estimation”.

avances en las técnicas de computación bayesiana, que permiten abordar de una forma relativamente cómoda los modelos.

La Encuesta Nacional de Demografía y Salud (ENDSA)<sup>2</sup> principal y posiblemente la única fuente de información para abordar ese tipo de análisis en el país, contiene el módulo del formulario utilizado para obtener la información que se refiere a la historia de nacimientos de cada una de las mujeres en edad fértil, información con la que se aborda el presente análisis. El descenso del nivel de mortalidad en los primeros años de vida (mortalidad neonatal, post-neonatal y post-infantil) suele realizarse graficando las estimaciones derivadas a partir de cada una de las ENDSAS. La forma o trayectoria de la mortalidad obtenida de esta forma, sin embargo, es distinta a la trayectoria obtenida con información de únicamente la última ENDSA. En el documento se presentan varias razones para encaminar el análisis de la evolución de la mortalidad con información de únicamente la última ENDSA, la realizada en el año 2008.

## 1. ANTECEDENTES

No se pudo encontrar en el ámbito nacional trabajos de investigación acerca del descenso de la mortalidad en los primeros años de vida y de sus determinantes, abordados con modelos de regresión desde un enfoque clásico o un enfoque bayesiano. Por esta razón, se podría decir que el presente estudio se constituye en una de las primeras experiencias en el país.

---

<sup>2</sup> En Bolivia se han ejecutado cinco encuestas nacionales de demografía y salud en el marco del Programa DHS, la primera en 1989 y la presente en 2008 (ENDSA 2008). El programa de Encuestas de Demografía y Salud se inició en 1984 como Programa DHS y desde fines de 1998 se conoce como Programa MEASURE DHS. El programa proporciona asistencia a instituciones gubernamentales y privadas en la implementación de encuestas nacionales en países en vías de desarrollo. Con financiamiento proveniente de la Agencia de los Estados Unidos para el Desarrollo Internacional (USAID), el Programa MEASURE DHS es implementado por Macro International Inc., con sede en Maryland, USA.

Fuera del país, no hubo estudios sobre el descenso de la mortalidad usando el modelo de regresión Poisson desde una perspectiva Bayesiana. En cambio, son varios los estudios llevados a cabo en relación al descenso de la mortalidad usando el modelo de regresión Poisson clásico, si bien esos estudios difieren en algunos aspectos del presente estudio. Uno de ellos se refiere a la “evolución de la mortalidad infantil, neonatal y postneonatal en Andalucía, 1975-1980” (Ramos y Nieto, 2003). El análisis se basó en información sobre defunciones anuales de los menores de un año provenientes de registros de estadísticas vitales. Mediante modelos de regresión de Poisson se estimaron los porcentajes de cambio anuales de las tasas. Otro estudio relacionado con la mortalidad, titulado “mortalidad por defectos del tubo neural en México, 1980-1997” (Ramirez y otros, 2003), tuvo el objetivo de describir la mortalidad en México por defectos del tubo neural, durante el periodo 1980-1997. La tendencia temporal fue evaluada por el porcentaje de cambio anual obtenido mediante un modelo de regresión de Poisson. En este estudio, las cifras de mortalidad por defectos del tubo neural también provienen de los registros de defunción. Se indica que la dispersión del modelo se valoró mediante el cociente entre el estadístico ji-cuadrado de Pearson y sus grados de libertad, utilizando un enfoque marginal que permitió que la varianza de las observaciones fuera diferente a la media. La literatura sobre el tema a la que se pudo acceder en general no expresa la forma en la que se abordó el problema de sobredispersión aparente o real.

## 2. PLANTEAMIENTO DEL PROBLEMA

La modelación de tasas de mortalidad implica una serie de complejidades que muchas veces las técnicas clásicas o frecuentistas se ven limitadas, en tal contexto urge una pronta actualización de metodologías que respalden los resultados que, si bien, responden de manera inmediata las interrogantes en el sector salud, a la vez no logran un análisis exhaustivo que desemboque en el

conocimiento y comprensión del descenso o evolución de las tasas de mortalidad en los primeros años de vida en Bolivia.

### **3. HIPÓTESIS**

La estimación de los parámetros desde un enfoque Bayesiano son mejores que los obtenidos a partir de un enfoque Clásico.

## **4. OBJETIVOS**

### **4.1 OBJETIVO GENERAL**

El objetivo general de la presente tesis es aplicar modelación Poisson, desde el punto de vista del enfoque clásico o frecuentista, como desde el punto de vista del enfoque bayesiano de la teoría estadística, para responder la siguiente interrogante: en Bolivia, en los últimos años, ¿en qué grupos de edad hubo mayor descenso de la mortalidad en los primeros años de vida?, y ¿cuáles son los principales factores que contribuyeron a ese descenso?, para establecer el descenso y los factores determinantes de la mortalidad en los primeros años de vida en Bolivia, durante el periodo 1993-2007.

### **4.2 OBJETIVOS ESPECÍFICOS**

- i) Determinar la tendencia de la mortalidad neonatal, post-neonatal y post-infantil durante el periodo de análisis. Es decir, establecer si la magnitud y el ritmo del descenso de la mortalidad es el mismo en todas las edades o, por el contrario, si ha sido más pronunciado en algunos tramos de edades.

- ii) Determinar el efecto de la educación y lugar de residencia sobre el descenso de la mortalidad neonatal, post-neonatal y post-infantil, durante el periodo de análisis.
- iii) Comparar los resultados de las estimativas de los parámetros de ambos enfoques (bayesiano y clásico).
- iv) Verificar la existencia de sobredispersión aparente o real.

## 5. JUSTIFICACIÓN

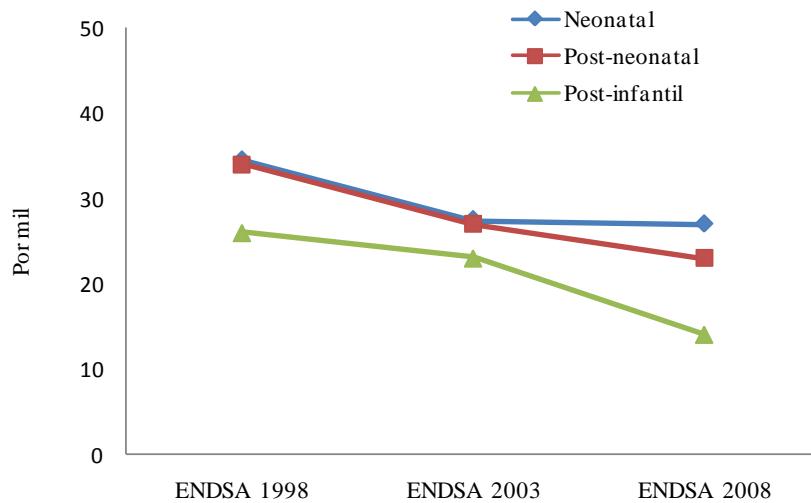
El enfoque clásico desde su existencia hasta la actualidad, ha sido de relevante progreso; sin embargo los métodos de estimación modernos exigen una constante actualización metódica dado la relevancia con la que se maneja los indicadores de salud y la importancia de obtener resultados interpretables del descenso del nivel de mortalidad en los primeros años de vida para evaluar la eficacia de las políticas y programas en el campo de la salud, implementados en el pasado. Pero no sólo para evaluar las acciones en este campo, también en otros campos, como en el de agricultura y de educación, debido a que la mortalidad es un fenómeno multicausal. Además, determinar y comprender las diferencias en cuanto al descenso o a la evolución de la mortalidad en los distintos grupos de población también es de importancia, pues proporciona elementos para adecuar o re-focalizar determinadas políticas. Una de las variables que tiene estrecha relación con la mortalidad es la edad. Como es sabido, el nivel de mortalidad difiere con la edad. La mortalidad en las primeras cuatro semanas de vida generalmente es mayor que en el tramo del segundo al onceavo mes de vida, y la mortalidad en éste tramo es mayor que en el tramo del segundo al cuarto año de vida.

Por otra parte, en el estudio del descenso de la mortalidad, la variable respuesta es una frecuencia. A tal hecho, la regresión Poisson basada en la distribución de probabilidad Poisson, es el método adecuado para modelar la frecuencia de muertes. Una limitación del modelo Poisson, sin embargo, es que la varianza es igual a la media, propiedad conocida como equidispersión. Pocos datos de frecuencia en la vida real son verdaderamente equidispersados. Empero, dependiendo de si la sobredispersión es aparente o real, hay varias maneras sugeridas en la literatura de poder corregirla o tratarla. Uno de los métodos adecuados, en caso de una real sobredispersión, es recurrir a una regresión Binomial negativa.

En Bolivia, cuando se trata de determinar el descenso de la mortalidad en los primeros cinco años de vida con base en los resultados de las diferentes Encuestas de Demografía y Salud, consideradas como las principales fuentes de información del País en este campo, surgen algunas contradicciones. El Gráfico 5.1 exhibe las tasas de mortalidad neonatal, post-neonatal y post-infantil obtenidas con las ENDSAS de 1998, 2003 y 2008. De acuerdo a estos resultados, la mortalidad neonatal habría descendido entre 1998 y 2003, y se habría mantenido aproximadamente constante entre 2003 y 2008.

Gráfico 5.1

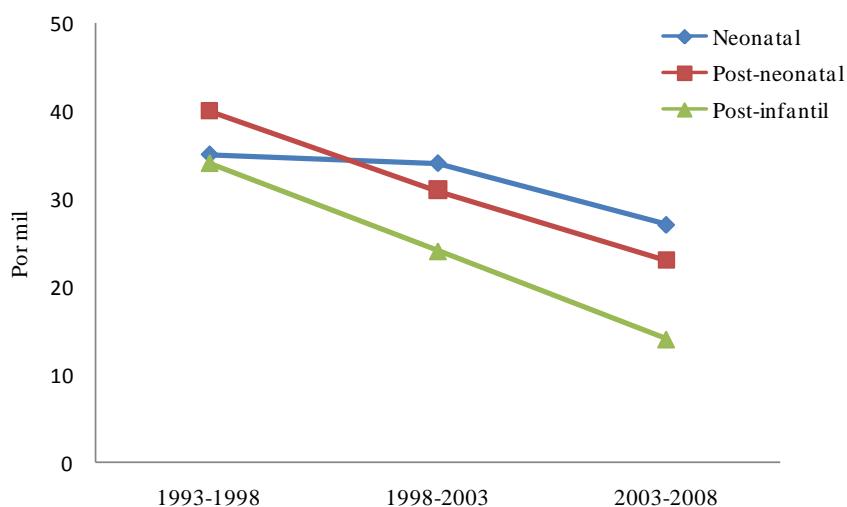
Bolivia: Evolución de las tasas de mortalidad neonatal, post-neonatal y post-infantil, combinando resultados de distintas ENDSAS



Sin embargo, cuando se trata de determinar la tendencia de la mortalidad a partir de la ENDSA más reciente, es decir la realizada en el año 2008, la tendencia es distinta. El Gráfico 5.2 muestra los resultados. Ahora, la tasa de mortalidad neonatal se habría mantenido más o menos constante en el periodo 1998-2003 y habría descendido en el siguiente periodo, 2003-2008. Éste es un resultado opuesto al anterior. También se puede observar algunas diferencias entre ambos gráficos con relación a la tendencia de la mortalidad post-neonatal y post-infantil.

Gráfico 5.2

Bolivia: Evolución de las tasas de mortalidad neonatal, post-neonatal y post-infantil, según la ENDSA 2008



Obviamente la pregunta inmediata es: ¿cuál de las dos formas de ver la tendencia es la correcta? ¿La tendencia tendría que ser deducida a partir de los resultados de las distintas ENDSAS, como en el Gráfico 5.1, o a partir de la ENDSA más reciente, como en el Gráfico 5.2? En el caso específico de la mortalidad en los primeros años de vida, es más razonable determinar la tendencia a partir de la ENDSA más reciente. Al menos hay cuatro razones para elegir esta opción. La primera, tiene que ver con los diferentes tamaños de muestra de las distintas ENDSAS. La cantidad de hogares entrevistados en la encuesta de 1998 (12.109 hogares) es inferior a la de las encuestas realizadas en los años 2003 y 2008 (19.207 y 19.564 hogares, respectivamente). Puesto que el tamaño de muestra influye en la precisión del estimador, y si la tendencia es derivada a partir de la comparación de estimadores basados en distintos tamaños de muestra, entonces la comparación de estimadores imprecisos con otros relativamente precisos puede distorsionar la tendencia. La segunda, tiene relación con las unidades primarias de muestreo. La muestra de las tres ENDSAS se la obtuvo en dos etapas; en la primera etapa se seleccionaron conglomerados de hogares denominadas unidades primarias de muestreo, y en

la segunda etapa se seleccionó un número de hogares dentro de cada unidad primaria. En consecuencia, puesto que las tres ENDSAS tienen distintas unidades primarias de muestreo, la tendencia de la mortalidad podría estar influenciada por estas diferencias muestrales. La tercera, tiene relación con las unidades muestreadas en la segunda etapa. Por la misma razón a la anterior, los hogares y las mujeres en edad fértil entrevistados en las tres ENDSAS no son los mismos. Este hecho podría también influir en la evaluación de la tendencia. La cuarta razón está vinculada con las diferencias de errores en la declaración de la información, por ejemplo en la declaración de las fechas de muerte de los hijos. Entonces, la determinación de la tendencia de la mortalidad en los primeros cinco años de vida a partir de la simple comparación de resultados de las distintas ENDSAS podría estar distorsionada por estas cuatro causas, entre otras. Por estas razones, el análisis de tendencia en la presente investigación se basa en los resultados de la ENDSA realizada en el año 2008.

## 6. ALCANCES Y LIMITACIONES

El estudio tiene alcance a nivel nacional para infantes desde los 0 meses de edad hasta los 59 meses enmarcados desde el año 1993 hasta el año 2007 con la limitante de no contar con información actual de los últimos cinco años. Adicionalmente, al modelar la información desde la perspectiva Bayesiana conduce a enfrentar limitaciones de tipo metodológico tales como las contantes actualizaciones metódicas que si bien, algunas refutan técnicas anteriormente expuestas, a la vez proponen nuevas alternativas que generalmente no están explícitas en los textos de consulta y mucho menos en programas tradicionales de modelación Bayesiana. De manera que, lo más recomendable es abordar en primera instancia la modelación Poisson desde el enfoque Clásico y después desde el enfoque Bayesiano.

## ENFOQUE CLÁSICO

### 7. LA FAMILIA EXPONENCIAL

Muchas funciones de densidad de probabilidad o funciones de masa de probabilidad pueden ser englobadas en una familia muy particular denominada familia exponencial. El concepto de familia exponencial fue introducido en la Estadística por Ronald Fisher (1934), quien desarrolló la idea de que estas distribuciones comúnmente aplicadas son realmente casos especiales de una clasificación más general que él denominó familia exponencial.

#### 7.1 DEFINICIÓN

Sea la variable aleatoria  $Y$  con función de densidad o función de masa de probabilidad  $f(y|\theta,\phi)$ , dependiente de los parámetros  $\theta$  y  $\phi$ . Esta familia de funciones de densidad o de masa de probabilidad, pertenece a la familia exponencial si puede ser escrita de la forma

$$f(y|\theta,\phi) = \exp \frac{(y\theta - b(\theta))}{a(\phi)} + c(y;\phi), \quad (7.1)$$

donde:

- $\theta$  recibe el nombre de parámetro canónico o parámetro natural,
- $\phi$  es denominado parámetro de dispersión,  $a(\phi)$  es una función dependiente del parámetro de dispersión,
- $b(\theta)$  es denominado la función cumulante, importante en el cálculo de los momentos de la distribución,
- $c(y,\phi)$  es una función que depende del parámetro de dispersión y de la información.

De tal manera (7.1) recibe el nombre de familia exponencial con dispersión. Esta distribución está expresada en su forma canónica debido a que  $\theta$  o bien es el mismo parámetro de la distribución de la variable aleatoria  $Y$  o bien es una transformación uno a uno del parámetro en la distribución de  $Y$ .

Cuando la función de densidad o función masa de probabilidad no tiene un parámetro de dispersión entonces  $a \phi = 1$ . En este caso, la expresión (7.1) se reduce a

$$f(y|\theta) = e^{y\theta - b(\theta) + c(y)},$$

la que también puede ser escrita como

$$\begin{aligned} f(y|\theta) &= e^{y\theta} e^{-b(\theta)} e^{c(y)} \\ &= a(\theta) b(y) e^{y\theta} \end{aligned}$$

donde  $a(\theta) = e^{-b(\theta)}$ ,  $b(y) = e^{c(y)}$ ,  $a(\theta) > 0$  y  $b(y) > 0$  para todo  $\theta$  e  $y$ , respectivamente.

### **Ejemplo 1: La distribución Poisson**

Sea  $\mu$  como la tasa de ocurrencia o el número esperado de veces que un evento ocurre en un periodo de tiempo dado. Se define la variable aleatoria  $Y$  como el número de veces que un evento ocurre. La relación entre  $\mu$  y  $y$  es especificada por la distribución Poisson como sigue

$$f(y|\mu) = P(Y=y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \quad y = 0, 1, 2, \dots$$

escrita en la forma de la familia exponencial de distribuciones (7.1)

$$f(y|\mu) = e^{y\ln(\mu)-\mu-\ln(y!)}, \quad (7.2)$$

con

$$\begin{aligned} \theta &= \ln(\mu) \\ b(\theta) &= \mu = e^\theta \\ c(y) &= -\ln(y!) \\ a(\phi) &= 1 \end{aligned}$$

## 7.2 MOMENTOS EN LA FAMILIA EXPONENCIAL

En el trabajo de derivación de los momentos para la familia exponencial de distribuciones, es de vital importancia conocer algunos conceptos de condiciones de regularidad.

### 7.2.1 Condiciones de Regularidad

Los resultados sobre consistencia y normalidad asintótica de los estimadores de máxima verosimilitud se cumplen si las condiciones de regularidad se cumplen<sup>3</sup>. Además, el estimador de máxima verosimilitud tiene la propiedad de alcanzar el límite inferior de Cramér-Rao y es por tanto completamente eficiente. El teorema del valor medio expresa mencionado resultado.

- $\theta \in \Theta$ , de lo contrario la función de verosimilitud no sería regular;

---

<sup>3</sup> Ver Casella, G., and R.L.Berger. 1990. "Statistical Inference".

- la información de Fisher es positiva y acotada, de lo contrario no sería una varianza;
- Se puede tomar derivadas (hasta de tercer orden) de bajo el signo integral,

$$\frac{\partial f(y|\theta)}{\partial \theta}, \quad \frac{\partial^2 f(y|\theta)}{\partial \theta^2}, \quad \frac{\partial^3 f(y|\theta)}{\partial \theta^3},$$

es decir el orden de derivada e integral se puede invertir;

$$\frac{\partial f(y|\theta)}{\partial \theta} dy, \quad \frac{\partial^2 f(y|\theta)}{\partial \theta^2} dy, \quad \frac{\partial^3 f(y|\theta)}{\partial \theta^3} dy.$$

Estas condiciones de regularidad se cumplen para las distribuciones de probabilidad pertenecientes a la familia exponencial.

### 7.2.2 Momentos

En función de los términos utilizados en (7.1) puede obtenerse expresiones generales para  $E(Y)$  y  $\text{Var}(Y)$ . Sea  $L$  la función log-verosimilitud de la familia exponencial, entonces de (7.1),

$$L(\theta; y) = \ln f(y; \theta),$$

muchos de los resultados claves acerca de los MLG relacionan para la derivada

$$U = \frac{dL}{d\theta},$$

donde  $U$  es llamado *score*.

Para encontrar los momentos de  $U$  usamos la identidad,

$$\frac{d \ln f(y; \theta)}{d\theta} = \frac{1}{f(y; \theta)} \frac{df(y; \theta)}{d\theta}, \quad (7.3)$$

para encontrar la esperanza de  $U$

$$E U = \frac{d \ln f(y; \theta)}{d\theta} f(y; \theta) dy = \frac{1}{f(y; \theta)} \frac{df(y; \theta)}{d\theta} f(y; \theta) dy = \frac{df(y; \theta)}{d\theta} dy,$$

donde la integración es sobre el dominio de  $y$ . Bajo ciertas condiciones de regularidad el último término del lado derecho expresado anteriormente es;

$$\frac{df(y; \theta)}{d\theta} dy = \frac{d}{d(\theta)} f(y; \theta) dy = \frac{d}{d(\theta)} 1 = 0$$

puesto que  $f(y; \theta) dy = 1$ . Entonces,

$$E U = 0. \quad (7.4)$$

Además si se diferencia (7.3) por segunda vez,

$$\frac{d}{d(\theta)} \frac{d \ln f(y; \theta)}{d\theta} f(y; \theta) dy = \frac{d^2}{d(\theta)^2} f(y; \theta) dy = \frac{d^2}{d(\theta)^2} 1 = 0,$$

el primer término del lado derecho es cero porque  $f(y; \theta) dy = 1$ , y el término del lado izquierdo se mantiene y es expresado como,

$$\frac{d^2 \ln f(y; \theta)}{d(\theta)^2} f(y; \theta) dy + \left( \frac{d \ln f(y; \theta)}{d\theta} \right) \frac{df(y; \theta)}{d\theta} dy.$$

Entonces, sustituyendo (7.3) en el segundo término obtenemos;

$$\frac{d^2 \ln f(y; \theta)}{d(\theta)^2} f(y; \theta) dy + \left( \frac{d \ln f(y; \theta)}{d\theta} \right)^2 f(y; \theta) dy = 0.$$

Por lo tanto

$$E - \frac{d^2 \ln f(y; \theta)}{d(\theta)^2} = E \left( \frac{d \ln f(y; \theta)}{d\theta} \right)^2$$

En términos del score estadístico, esto es;

$$E - U' = E(U^2)$$

donde  $U'$  denota la derivada de  $U$  con respecto a  $\theta$ . Desde  $E U = 0$ , la  $V U$  llamada información; es

$$V U = E U^2 = E - U'. \quad (7.5)$$

Se tiene

$$U = \frac{dL}{d\theta} = \frac{y - b' \theta}{a \phi}$$

$$\frac{dU}{d\theta} = \frac{\partial^2 L}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}$$

$$V U = E - \frac{dU}{d\theta} = \frac{b''(\theta)}{a(\phi)}.$$

Con ambos resultados, (7.4) y (7.5), es posible generalizar situaciones con más de un parámetro.

Por un argumento análogo usado para derivar (7.5) en el caso de una única variable aleatoria con un único parámetro, se verifica la siguiente igualdad

$$E \left[ \frac{\partial L}{\partial \theta_j} \frac{\partial L}{\partial \theta_k} \right] = E \left[ -\frac{\partial^2 L}{\partial \theta_j \partial \theta_k} \right].$$

Luego, la matriz de información<sup>4</sup> también puede ser expresada como

$$\Sigma = \begin{matrix} & \vdots & \cdots & \vdots \\ \vdots & E \left[ -\frac{\partial^2 L}{\partial \theta_j \partial \theta_k} \right] & \vdots \\ & \vdots & \cdots & \vdots \end{matrix}. \quad (7.6)$$

Los resultados (7.4), (7.5) y sus correspondientes generalizaciones se cumplen para distribuciones de la familia exponencial bajo las condiciones de regularidad, así se procede a obtener las expresiones generales para  $E Y_i$  y  $V Y_i$  bajo muestreo aleatorio, entonces por (7.4)

$$E \left( \frac{\partial L_i}{\partial \theta_i} \right) = \frac{E Y_i - b' \theta_i}{a(\phi)} = 0;$$

es decir,

$$\mu_i = E Y_i = b'(\theta_i). \quad (7.7)$$

De (7.5),

---

<sup>4</sup>  $\Sigma$  en todo el documento expresa la matriz de información esperada (o de Fisher).

$$-E \frac{\partial^2 L_i}{\partial \theta_i^2} = -E -\frac{b''(\theta_i)}{a(\phi)} = \frac{b''(\theta_i)}{a(\phi)} = E \frac{\partial L_i}{\partial \theta_i}^2 = E \frac{Y_i b' \theta_i}{a \phi}^2,$$

$$\frac{b''(\theta_i)}{a(\phi)} = E \left[ \frac{Y_i - b' \theta_i}{a \phi} \right]^2 = \frac{Var(Y_i)}{[a \phi]^2},$$

de modo que

$$Var Y_i = b'' \theta_i a \phi = V(\mu_i) a \phi, \quad (7.8)$$

donde  $V(\mu_i) = b'' \theta_i$  se denomina función varianza. En suma, con base en (7.7) y (7.8) se puede ver que la función  $b$  de la expresión (7.1) determina los momentos de  $Y_i$ .

### **Ejemplo 2: La distribución Poisson**

En el ejemplo 1 se dedujo, expresado en términos de la observación  $i$ , que

$$\begin{aligned} \theta_i &= \ln(\mu_i), \\ b \theta_i &= \mu_i = e^{\theta_i}, \\ a \phi &= 1. \end{aligned}$$

Luego, el valor esperado y la varianza de la distribución Poisson, usando (7.7) y (7.8), quedan formuladas como

$$\begin{aligned} E Y_i &= b' \theta_i = e^{\theta_i} = \mu_i \\ Var Y_i &= b'' \theta_i a \phi = e^{\theta_i} = \mu_i \end{aligned}$$

### 7.2.3 Sobredispersión

Se asume debido a la naturaleza de la distribución Poisson, que  $Var(Y) = \sigma^2 E(Y)$ ; donde  $\sigma^2$  es el parámetro de dispersión y se asume constante. Es decir la distribución de Poisson se caracteriza por la equidispersión, esto es:  $Var(Y) = E(Y)$ .

Sin embargo, un problema que se da con cierta frecuencia en este modelo es que la relación media-varianza no es equitativa. Las desviaciones en relación a la equidispersión pueden resultar en:

- Sobredispersión: Si  $Var(Y) > E(Y)$ , es decir si  $\sigma^2 > 1$ .
- Infradispersión o Subdispersión:  $Var(Y) < E(Y)$ , es decir si  $\sigma^2 < 1$ .

Cuando existe exceso de variación en los datos, las estimaciones de los errores estándar pueden resultar sesgadas, pudiendo presentarse errores en las inferencias a partir de los parámetros del modelo de regresión.

## 8. MODELOS LINEALES GENERALIZADOS

Nelder y Wedderburn introdujeron la teoría de los modelos lineales generalizados en 1972, como su nombre indica los modelos lineales generalizados generalizan el modelo de regresión lineal permitiendo utilizar el mismo tipo de modelización, especificación, estimación y diagnóstico, aquí radica su atractivo, para variables dependientes no normales como por ejemplo variables contadoras (distribución de Poisson); variables dicotómicas (distribución Binomial), etc.

Los modelos GLM son miembros de la familia exponencial con función de densidad de la forma de (7.1)

### 8.1 COMPONENTES DE UN MODELO LINEAL GENERALIZADO (MLG)

En definitiva, un modelo lineal generalizado tiene tres componentes: (1) un componente aleatorio o variable respuesta (2) un componente sistemático y (3) una función de enlace que vincula los componentes aleatorio y sistemático.

1. La variable repuesta o componente aleatorio  $Y$  con observaciones independientes  $y_1, y_2, \dots, y_n$  con distribución perteneciente a la familia exponencial.
2. El componente sistemático de un MLG, se refiere al conjunto de parámetros  $\beta$  y variables explicativas en una función de predicción lineal.

$$\eta_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, 2, \dots, n \quad (8.1)$$

Esta combinación lineal de variables explicativas se denomina predictor lineal, matricialmente la fórmula (8.1) puede ser expresada para las n observaciones como

$$\eta = X\beta,$$

donde el vector  $\eta$  es de dimensión  $n \times 1$ , la matriz  $X$  de valores de las variables explicativas es de dimensión  $n \times p$  y el vector de parámetros  $\beta$  de dimensión  $p \times 1$ .

3. Una función enlace  $g$  que conecta los componentes aleatorio y sistemático, es monótona y diferenciable tal que

$$\eta_i = g(\mu_i) = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, 2, \dots, n \quad (8.2)$$

con

$$\mu_i = E(y_i)$$

En resumen, un MLG es un modelo lineal para la transformación de la media de una variable respuesta que tiene distribución en la familia exponencial.

### **Ejemplo 3: El MLG Poisson**

En un MLG Poisson, el componente aleatorio se conecta con el componente sistemático mediante la función de enlace canónica  $\ln$ ; es decir

$$\eta_i = \ln \mu_i$$

$$\ln \mu_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n \quad (8.3)$$

$$\mu_i = e^{\sum_{j=1}^p \beta_j x_{ij}}, \quad i = 1, \dots, n \quad (8.4)$$

y en términos matriciales,

$$\mu = e^{X\beta}, \quad (8.5)$$

donde  $\mu$  es un vector de dimensión  $n \times 1$ .

## 8.2 ECUACIONES DE VEROSIMILITUD PARA EL MLG

Dadas las observaciones  $Y_1 = y_1, \dots, Y_n = y_n$ , la función de verosimilitud es

$$L(\theta, \phi, y) = \prod_{i=1}^n f(y_i; \theta_i, \phi) = \exp \left( \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^n c(y_i, \phi) \right)$$

donde  $y = (y_1, \dots, y_n)$  y  $\theta = (\theta_1, \dots, \theta_n)$  es el parámetro canónico.

La función de log – verosimilitud es

$$l(\theta, \phi, y) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^n c(y_i, \phi) = \sum_{i=1}^n l_i(\theta_i, \phi, y_i) \quad (8.7)$$

y se obtiene al resolver las ecuaciones de verosimilitud

$$U_j = \frac{\partial l(\theta, \phi, y)}{\partial \beta_j} = 0, \quad j = 1, \dots, p;$$

se verifica la relación de dependencia

$$l_i \leftarrow \theta_i \leftarrow \mu_i \leftarrow \eta_i \leftarrow \beta,$$

donde  $\theta_i \leftarrow \mu_i$  significa que  $\theta_i$  es función de  $\mu_i$ . Puesto que  $\mu_i = b'(\theta_i)$ , las relaciones son

- $l_i = l_i(\theta_i) = \frac{y_i \theta_i - b(\theta_i)}{a_i \phi} + c(y_i, \phi)$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i \theta_i - b'(\theta_i)}{a_i \phi} = \frac{y_i - \mu_i}{a_i \phi}$$

- $\mu_i = \mu_i(\theta_i) = b'(\theta_i)$

$$\theta_i = \theta_i(\mu_i) = b'^{-1}(\mu_i)$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{b''(\theta_i)}$$

- $\eta_i = \eta_i(\mu_i) = g(\mu_i)$

$$\mu_i = \mu_i(\eta_i) = g^{-1}(\eta_i)$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{g'(\mu_i)}$$

- $\eta_i = \eta_i(\beta) = x_i^t \beta$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Al aplicar la regla de la cadena y la igualdad  $V(Y_i) = a_i \phi b'' \theta_i$ , se deduce que

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{y_i - \mu_i x_{ij}}{V(Y_i) g'(\mu_i)} \quad (8.8)$$

Finalmente, sustituyendo en las ecuaciones de verosimilitud,

$$\sum_{i=1}^n \frac{\partial}{\partial \beta_j} l_i \theta_i, \phi, y_i = 0, \quad j = 1, \dots, p \quad (8.9)$$

se obtiene el resultado enunciado; es decir

$$\sum_{i=1}^n \frac{y_i - \mu_i x_{ij}}{a_i \phi b'' \theta_i g'(\mu_i)} = \sum_{i=1}^n \frac{y_i - \mu_i x_{ij}}{V(Y_i) g'(\mu_i)} = 0, \quad j = 1, \dots, p \quad (8.10)$$

Aunque  $\beta$  no aparece en las ecuaciones, está implícito a través de  $\mu_i$ , puesto que  $\mu_i = g^{-1}(\sum_{j=1}^p \beta_j x_{ij})$ . Debe notarse que diferentes funciones de enlace producen diferentes conjuntos de ecuaciones<sup>5</sup>.

#### **Ejemplo 4: El MLG Poisson**

El modelo log-lineal Poisson tiene la forma

$$\ln \mu_i = \sum_{j=1}^p \beta_j x_{ij}, \quad i = 1, \dots, n,$$

que en términos matriciales se lo puede expresar como

$$\ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}.$$

---

<sup>5</sup> Ver Dobson, A.J. 1990. "An Introduction to Generalized Linear Models".

Para el enlace canónico  $\ln, \eta_i = \ln \mu_i$ ; es decir  $\mu_i = e^{\eta_i}$ . Luego,

$$\frac{\partial \mu_i}{\partial \eta_i} = e^{\eta_i} = \mu_i. \quad (8.11)$$

Además, como  $\text{Var } Y_i = \mu_i$ , las ecuaciones de verosimilitud (8.11) se convierten en

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial L_i}{\partial \beta_j} = \sum_{i=1}^n y_i - \mu_i x_{ij} = 0, \quad j = 1, \dots, p \quad (8.12)$$

### 8.3 EVALUACIÓN DE AJUSTE EN UN MLG

El proceso de ajustar un modelo a un conjunto de datos puede ser considerado como una forma de remplazar el conjunto de valores de los datos  $y$  por un conjunto de valores ajustados  $\mu$  derivados del modelo generalmente con un reducido número de parámetros. En general los valores de  $\mu$  no son iguales a los valores de  $y$ ; entonces la pregunta es cuánta discrepancia hay entre ellos (McCullagh y Nelder, 1989). Hay varias medidas de discrepancia o bondad de ajuste, pero en este apartado se considera aquélla derivada del logaritmo de una razón de verosimilitudes, denominada devianza.

#### 8.3.1 La función Desvío

Existen varias medidas para verificar la bondad de ajuste, Nelder y Wedderburn (1972) propusieron como medida de discrepancia a la función desvío o devianza, cuyo objetivo principal es la verificación de la adecuación del modelo en cuanto a las discrepancias locales que, en caso de ser significativas, pueden implicar el desarrollo de un nuevo modelo.

Se sabe que ajustar un modelo estadístico a un conjunto de datos, es resumir razonablemente la información de  $n$  observaciones para  $p$  parámetros, o sea, es sustituir un conjunto de valores observados  $y$  por un conjunto de valores ajustados  $\mu$ , con un número menor de parámetros. Por ejemplo, un modelo más simple, llamado modelo nulo (modelo corriente), que contiene apenas un parámetro que representa a la media  $\mu$  común en todas las observaciones  $y$ 's. Por otro lado, un modelo saturado contiene  $n$  parámetros, uno para cada observación.

En la práctica, se procura un modelo con  $p$  parámetros situado entre esos dos límites, ya que un modelo corriente (modelo nulo) es mucho más simple en cuanto un modelo saturado es no – informativo. Por ejemplo, un modelo saturado es útil para medir la discrepancia de un modelo intermedio en investigación con  $p$  parámetros ( $p < n$ ).

La función desvío viene expresada como:

$$S_p = 2(l_n - l_p); \quad (8.13)$$

Siendo  $l_n$  y  $l_p$  los máximos de la función de verosimilitud para los modelos saturado y corriente, asimilando que un modelo saturado es usado como base de medida de ajuste de un modelo corriente.

Si  $Y_1, \dots, Y_n$  es una muestra aleatoria con distribución perteneciente a la familia exponencial (7.1). Sea  $\theta_i = \theta(\mu_i)$  y  $\theta_i = \theta(y_i)$  las estimaciones máximo verosímil de los parámetros canónicos para un modelo saturado y corriente, respectivamente; entonces se tiene

$$l_p = \prod_{i=1}^n l(\theta_i, \phi; y_i) = \prod_{i=1}^n \frac{(y_i \theta_i - b(\theta_i))}{a_i(\phi)} + c(y_i; \phi)$$

Un log – verosímil maximizado sobre  $\beta$  para un  $\phi$  fijo. Sea

$$l_n = \prod_{i=1}^n l(\theta_i, \phi; y_i) = \prod_{i=1}^n \frac{(y_i \theta_i - b(\theta_i))}{a_i(\phi)} + c(y_i; \phi)$$

Un log – verosímil para un modelo saturado con  $n$  parámetros. Asumiendo que  $a_i(\phi) = \frac{\phi}{\lambda_i}$ , se puede escribir

$$2(l_n - l_p) = \frac{\sum_{i=1}^n 2\lambda_i y_i \theta_i - b(\theta_i) + b(\theta_i)}{\phi} = \frac{D(y; \mu)}{\phi} = \frac{D}{\phi} \quad (8.14)$$

donde  $S_p$  y  $D$  son denominados desvío escalonado y desvío, respectivamente. El desvío es apenas función de los datos  $y$  y de las medias ajustadas  $\mu$ . El desvío escalonado  $S_p$  depende de  $D$  y del parámetro de dispersión  $\phi$ . Así la el desvío se define como el desvío escalonado por el parámetro de dispersión:

$$D = D(y; \mu) = \sum_{i=1}^n 2\lambda_i y_i \theta_i - b(\theta_i) + b(\theta_i) = S_p \phi \quad (8.15)$$

### **Ejemplo 5: El MLG Poisson**

La distribución Poisson (7.2), expresada para la observación  $i$ , es

$$f(y_i | \mu_i) = e^{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)}.$$

Si  $Y_1, \dots, Y_n$  es una muestra aleatoria de la Poisson, su distribución conjunta está dada por

$$f(\mathbf{y} | \boldsymbol{\mu}) = f(y_1, \dots, y_n | \mu_1, \dots, \mu_n) = \prod_{i=1}^n f(y_i | \mu_i) = e^{-\sum_{i=1}^n y_i \ln(\mu_i) - \sum_{i=1}^n \mu_i - \ln \Gamma(y_i + 1)},$$

con función log-verosimilitud

$$L(\boldsymbol{\mu} | \mathbf{y}) = \sum_{i=1}^n y_i \ln(\mu_i) - \sum_{i=1}^n \mu_i - \ln \Gamma(y_i + 1). \quad (8.16)$$

Como  $\ln \mu_i = \mathbf{X}_i \boldsymbol{\beta}$  y  $\mu_i = e^{x_i \boldsymbol{\beta}}$ , entonces la función log-verosimilitud puede ser reescrita como

$$= \sum_{i=1}^n y_i (\mathbf{X}_i \boldsymbol{\beta}) - e^{x_i \boldsymbol{\beta}} - \ln \Gamma(y_i + 1). \quad (8.17)$$

La función desvío o devianza requerida para el algoritmo del MLG, está definida como

$$\begin{aligned} S_p &= 2(l_n - l_p) \\ &= 2 \sum_{i=1}^n y_i \ln(y_i) - y_i - \ln \Gamma(y_i + 1) - 2 \sum_{i=1}^n y_i \ln(\mu_i) - \mu_i - \\ &\quad \ln \Gamma(y_i + 1), \\ &= 2 \sum_{i=1}^n y_i \ln y_i - \ln(\mu_i) - y_i + \mu_i. \end{aligned} \quad (8.18)$$

$$= 2 \sum_{i=1}^n y_i \ln \frac{y_i}{\mu_i} - y_i + \mu_i \quad (8.19)$$

El concepto de devianza también es útil en la comparación de dos modelos, consideremos dos modelos con la misma distribución y función enlace. Sean

$M_1 = (Y_1, \beta_1)$  y  $M = (Y, \beta)$  las matrices de diseño y los vectores de parámetros de ambos modelos. Se dice que  $M_1 \subset M$  (el modelo  $M_1$  está anidado en el modelo  $M$ ) si y sólo si:

$$Y \ n \times p = X_1 \ n \times p_1, X_2 \ n \times p_2$$

$$\beta^t \ 1 \times p = \beta_1^t \ 1 \times p_1, \beta_2^t \ 1 \times p_2, \text{ con } p_1 + p_2 = p, 0 < p_1, p_2 < p.$$

Sea  $\beta^t = (\beta_1^t, \beta_2^t)$ , con  $\dim \beta_1 = p_1 < p = \dim(\beta)$ . Se trata de contrastar

$$\begin{aligned} H_0: \beta_2 &= 0 \\ H_1: \beta_2 &\neq 0 \end{aligned}$$

Limitándose al caso  $a_i \phi = a_i \phi$  y se considera dos situaciones:

- a)**  $\phi$  conocido
- b)**  $\phi$  desconocido

Sean  $D_1$  y  $D$  los desvíos (no escalados) asociados a los modelos  $M_1$  y  $M$  respectivamente.

- a)** Si  $\phi$  es conocido. Bajo  $H_0$ , se verifica que

$$T_1 = \frac{D_1 - D}{\phi} \sim \chi_{p-p_1}^2$$

En consecuencia, se rechaza  $H_0$  si  $T_1 > \chi_{p-p_1}^2, 1 - \alpha$

- b)** Si  $\phi$  es desconocido, se estima con  $\hat{\phi} = \frac{D}{n-p}$  bajo  $H_0$ , se verifica que

$$T_2 = \frac{D_1 - D}{\phi} = \frac{(D_1 - D)/(p - p_1))}{D/(n - p)} \sim F_{p-p_1, n-p}$$

En consecuencia, se rechaza  $H_0$  si  $T_2 > F_{p-p_1, n-p}, 1 - \alpha$ .

El diagnóstico de sobredispersión está basado en una prueba de Razón de Verosimilitud (RV). Esta consiste en probar la hipótesis de que el parámetro  $k = 0$ , donde  $k$  es el factor por el cual se incrementa la varianza de la distribución Poisson.

Según Cameron y Trivedi (1998) esta prueba tiene una distribución asintótica  $\chi^2_{(1-2\alpha, 1)}$ . Por tanto, rechazaremos  $H_0$  si la estadística es mayor que  $\chi^2_{(1-2\alpha, 1)}$ .

### 8.3.2 Análisis de Residuos

Cuando se realiza la evaluación del modelo, los residuos miden la discrepancia entre los valores observados de la variable dependiente y los valores ajustados para cada observación. Ellos pueden ser usados para detectar la mala especificación del modelo; para detectar observaciones con un pobre ajuste; y para detectar las observaciones influyentes, u observaciones que afectan los coeficientes estimados.

#### **Residuo de Devianza**

La devianza es una estadística importante en la derivación del MLG y en el proceso inferencial de los resultados. Si la distribución de la variable respuesta pertenece a la familia exponencial entonces se puede usar la *devianza residual*, definida como

$$d_i = signo(y_i - \mu_i) \sqrt{2 L(y_i | \mu_i) - L(\mu_i | y_i)}, \quad (8.20)$$

donde  $L(\mu_i | y_i)$  es la log-verosimilitud de  $y_i$  evaluada en  $\mu_i = \mu_i$  y  $L(y_i | \mu_i)$  es la log-verosimilitud evaluada en  $\mu_i = y_i$ . Una motivación para la devianza residual es que la suma de cuadrados de estos residuales es la estadística devianza; es decir  $D = 2 L(y | \mu) - L(\mu | y) = \sum_{i=1}^n d_i^2$ .

Sus propiedades distribucionales son las que hacen a que los investigadores elijan en primera instancia a este residuo<sup>6</sup>.

### **Residuo de Pearson**

Para datos de conteo no hay un residuo netamente con comportamiento normal. Este hecho conduce a muchos diferentes residuos. La corrección obvia por heterocedasticidad es el *residuo de Pearson*

$$p_i = \frac{(y_i - \mu_i)}{\sqrt{Var(Y_i)}}. \quad (8.21)$$

En muestras grandes este residuo tiene media cero y es homocedástico (con varianza uno), pero su distribución es asimétrica (Cameron y Trivedi, 1998).

### **Residuo de Anscombe**

A este residuo se lo identifica como la transformación de  $y$  que es más próximo a la normalidad, con media cero y varianza 1. Esta trasformación se la obtuvo para distribuciones de la familia exponencial y se la define como

---

<sup>6</sup> Ver McCullagh, P., and J.A. Nelder. 1989. "Generalized Linear Models".

$$r_i^A = \frac{A(y_i - A(\mu_i))}{A'(\mu_i) \sqrt{V(\mu_i)}}, \quad (8.22)$$

donde

$$A . = \frac{d\mu}{Var(Y)^{1/3}}.$$

La elección de la función  $A .$  fue hecha de modo que los residuos resultantes sean tan normales como sea posible.

### **Ejemplo 6: El MLG Poisson**

Para el modelo lineal generalizado Poisson se presenta los siguientes residuos:

#### *Residuo Devianza*

$$d_i = signo(y_i - \mu_i) \sqrt{2 y_i \ln \frac{y_i}{\mu_i} - y_i + \mu_i}. \quad (8.23)$$

#### *Residuo de Pearson*

$$p_i = \frac{y_i - \mu_i}{\mu_i}.$$

#### *Residuo de Anscombe*

$$A . = \frac{d\mu}{Var(Y)^{1/3}} = \frac{d\mu}{\mu^{1/3}} = \frac{3}{2} \mu^{2/3},$$

$$r_i^A = \frac{A(y_i - A(\mu_i))}{A'(\mu_i) \sqrt{V(\mu_i)}} = \frac{\frac{3}{2} y_i^{2/3} - \frac{3}{2} \mu_i^{2/3}}{\frac{3}{2} \frac{2}{3} \mu_i^{-1/3} \mu_i^{1/2}},$$

$$r_i^A = \frac{\frac{3}{2} y_i^{2/3} - \mu_i^{2/3}}{\mu_i^{1/6}}.$$

## 8.4 ESTIMACIÓN DE PARÁMETROS EN UN MLG

Este acápite da una pauta de cómo encontrar los estimadores de máxima verosimilitud  $\beta$  de los parámetros del MLG. Generalmente las ecuaciones de verosimilitud son no-lineales en  $\beta$ . A continuación se describe un método iterativo para resolver ecuaciones no-lineales y se desarrolla de dos maneras para determinar el máximo de una función de verosimilitud.

### 8.4.1 Método de Newton-Raphson

Este método utiliza la ecuación recurrente

$$\beta^{(r)} = \beta^{(r-1)} - H^{-1} \beta^{(r-1)} U \beta^{(r-1)}, \quad (8.24)$$

donde

$$U = \frac{\partial l}{\partial \beta_1}, \dots, \frac{\partial l}{\partial \beta_p}^t, \quad \beta^{(r)} = \beta_1^{(r)}, \dots, \beta_p^{(r)}^t, \quad H = \frac{\partial^2 l}{\partial \beta_j \partial \beta_k}_{j,k=1,\dots,p}$$

- $\beta^{(r)}$  es el valor estimado de  $\beta$  en la  $r$ -ésima iteración del algoritmo.
- $H^{-1} \beta^{(r-1)}, U \beta^{(r-1)}$  son  $H^{-1}$  y  $U$  evaluadas en  $\beta^{(r-1)}$ .

Justificación. El algoritmo de Newton – Raphson se apoya en el desarrollo en serie de Taylor. Si  $\beta^*$  es solución de las ecuaciones de verosimilitud; es decir,

$$U(\beta^*) = 0,$$

y  $\beta^{(0)}$  es un valor arbitrario de  $\beta$ , entonces el desarrollo de Taylor de primer orden garantiza la siguiente aproximación

$$0 = U(\beta^*) \cong U(\beta^0) + H(\beta^0)(\beta^* - \beta^0)$$

donde  $H = \frac{\partial U}{\partial \beta} = \frac{\partial^2 l}{\partial \beta_j \partial \beta_k}$  es la matriz Hessiana. De la anterior ecuación se obtiene

$$\beta^* \cong \beta^0 - H^{-1}(\beta^0)U(\beta^0),$$

que sirve de base para plantear la ecuación recurrente.

De (8.10), la gradiente, es decir el vector de derivadas de primer orden, queda expresada como

$$L' \boldsymbol{\beta} = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{matrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p} \end{matrix} = \begin{matrix} n \frac{(y_i - \mu_i)x_{i1}}{a \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} \\ \vdots \\ n \frac{(y_i - \mu_i)x_{ij}}{a \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} \\ \vdots \\ n \frac{(y_i - \mu_i)x_{ip}}{a \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} \end{matrix},$$

(8.25)

evaluada en  $\boldsymbol{\beta}^{r-1}$ .

Por otra parte, el elemento  $(j, k)$  en la matriz Hessiana puede obtenerse considerando que

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}.$$

$$\text{Como } L(\boldsymbol{\beta}) = \sum_{i=1}^n L_i, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial L_i}{\partial \beta_j}.$$

Luego,

$$\begin{aligned} \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} &= \frac{\partial}{\partial \beta_k} \sum_{i=1}^n \frac{\partial L_i}{\partial \beta_j}, \\ &= \sum_{i=1}^n \frac{\partial}{\partial \beta_k} \frac{\partial L_i}{\partial \beta_j}, \\ &= \sum_{i=1}^n \frac{\partial}{\partial \beta_k} \frac{(y_i - \mu_i)x_{ij}}{a \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}. \end{aligned} \quad (8.26)$$

Pero,

$$\frac{\partial}{\partial \beta_k} \frac{(y_i - \mu_i)x_{ij}}{a \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = \frac{x_{ij}}{a \phi} \frac{\partial}{\partial \beta_k} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}, \quad (8.27)$$

donde

$$\begin{aligned} \frac{\partial}{\partial \beta_k} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial}{\partial \beta_k} \frac{(y_i - \mu_i)}{V(\mu_i)} + \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial}{\partial \beta_k} \frac{\partial \mu_i}{\partial \eta_i}, \\ &= \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial}{\partial \mu_i} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} + \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial}{\partial \eta_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k}, \\ &= \frac{\partial \mu_i}{\partial \eta_i} \frac{-V \mu_i - (y_i - \mu_i) \frac{\partial V(\mu_i)}{\partial \mu_i}}{V^2(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ik} + \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial^2 \mu_i}{\partial \eta_i^2} x_{ik}, \\ &= \frac{\partial \mu_i}{\partial \eta_i}^2 - \frac{1}{V \mu_i} - \frac{(y_i - \mu_i)}{V^2(\mu_i)} \frac{\partial V(\mu_i)}{\partial \mu_i} x_{ik} + \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial^2 \mu_i}{\partial \eta_i^2} x_{ik}, \end{aligned}$$

$$\begin{aligned}
 &= -\frac{1}{V \mu_i} \frac{\partial \mu_i}{\partial \eta_i}^2 - (y_i \\
 &- \mu_i) \frac{1}{V^2(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}^2 \frac{\partial V(\mu_i)}{\partial \mu_i} - \frac{1}{V \mu_i} \frac{\partial^2 \mu_i}{\partial \eta_i^2} x_{ik}, \\
 &= -\frac{1}{V \mu_i} \frac{\partial \mu_i}{\partial \eta_i}^2 - (\mu_i - y_i) \frac{1}{V^2(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}^2 \frac{\partial V(\mu_i)}{\partial \mu_i} - \frac{1}{V \mu_i} \frac{\partial^2 \mu_i}{\partial \eta_i^2} x_{ik}.
 \end{aligned}$$

Remplazando en (8.24) la última expresión se tiene

$$\begin{aligned}
 \frac{\partial}{\partial \beta_k} \frac{(y_i - \mu_i)x_{ij}}{a \phi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} &= -\frac{1}{a \phi} \frac{1}{V \mu_i} \frac{\partial \mu_i}{\partial \eta_i}^2 - (\mu_i - y_i) \frac{1}{V^2(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}^2 \frac{\partial V(\mu_i)}{\partial \mu_i} - \\
 &\quad \frac{1}{V \mu_i} \frac{\partial^2 \mu_i}{\partial \eta_i^2} x_{ij} x_{ik},
 \end{aligned}$$

lo que su vez, sustituyendo en (8.24) produce

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^n \frac{1}{a \phi} \frac{1}{V \mu_i} \frac{\partial \mu_i}{\partial \eta_i}^2 - (\mu_i - y_i) \frac{1}{V^2(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}^2 \frac{\partial V(\mu_i)}{\partial \mu_i} - \frac{1}{V \mu_i} \frac{\partial^2 \mu_i}{\partial \eta_i^2} x_{ij} x_{ik}. \quad (8.28)$$

Una vez lograda la optimización, es importante considerar a la matriz Hessiana observada para estimar la matriz de covarianzas de  $\beta$ .

### Ejemplo 7: El MLG Poisson

El vector gradiente  $\frac{\partial L}{\partial \beta}$  queda expresado como

$$L' \boldsymbol{\beta} = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{matrix} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} \\ \vdots \\ \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p} \end{matrix} = \begin{matrix} \mathbf{y} - \mathbf{u}' \mathbf{x}_1 \\ \vdots \\ \mathbf{y} - \mathbf{u}' \mathbf{x}_j \\ \vdots \\ \mathbf{y} - \mathbf{u}' \mathbf{x}_p \end{matrix},$$

Por otra parte, la matriz Hessiana observada puede ser escrita como

$$H = \begin{matrix} \vdots & \cdots & \vdots \\ \vdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} & \vdots \\ \vdots & \cdots & \vdots \end{matrix} \quad (8.29)$$

$$H = - \begin{matrix} \mathbf{x}_1'[\mathbf{x}_1 \mathbf{u}] & \cdots & \mathbf{x}_1'[\mathbf{x}_k \mathbf{u}] & \cdots & \mathbf{x}_1'[\mathbf{x}_p \mathbf{u}] \\ \vdots & & \vdots & & \vdots \\ \mathbf{x}_j'[\mathbf{x}_1 \mathbf{u}] & \vdots & \mathbf{x}_j'[\mathbf{x}_k \mathbf{u}] & \vdots & \mathbf{x}_j'[\mathbf{x}_p \mathbf{u}] \\ \vdots & & \vdots & & \vdots \\ \mathbf{x}_p'[\mathbf{x}_1 \mathbf{u}] & \cdots & \mathbf{x}_p'[\mathbf{x}_k \mathbf{u}] & \cdots & \mathbf{x}_p'[\mathbf{x}_p \mathbf{u}] \end{matrix}.$$

Con matriz de covarianzas estimada de  $\boldsymbol{\beta}$

$$Cov \boldsymbol{\beta} = - H^{-1}, \quad (8.30)$$

### 8.4.2 Método de Fisher Scoring

Este método emplea la ecuación recurrente

$$\beta^{(r)} = \beta^{(r-1)} + \mathfrak{U}^{-1} \beta^{(r-1)} U \beta^{(r-1)}, \quad (8.31)$$

donde

$$\mathfrak{U} = -E[H] = \sum_{i=1}^n E - \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \Big|_{j,k=1,\dots,p}$$

### 8.4.3 Mínimos Cuadrados Iterativamente Reponderados (MCIR)

El método de puntuaciones de Fisher se puede expresar de la siguiente forma

$$\beta^{(r)} = (X^t W^{-1} (\beta^{(r-1)}) X)^{-1} X^t W^{-1} (\beta^{(r-1)}) Z \beta^{(r-1)}$$

donde

$$W = \text{diag } V[Y_i g'(\mu_i)^2]; \quad i = 1, \dots, n$$

$$Z = X\beta + \text{diag } g' \mu_1, \dots, g' \mu_n \quad y - \mu.$$

Dada la similitud de las dos últimas ecuaciones, el método de puntuaciones de Fisher también recibe el nombre de “algoritmo de mínimos cuadrados iterativamente re ponderados”.

Para llegar a la última ecuación, se empieza por multiplicar la ecuación (8.30) por  $\mathfrak{U}(\beta^{(r-1)})$ .

Entonces

$$\mathfrak{U} \beta^{(r-1)} \beta^r = \mathfrak{U} \beta^{(r-1)} \beta^{(r-1)} + U \beta^{(r-1)} \quad (8.32)$$

Al calcular las expresiones de los términos de  $\nabla \beta$ . Se obtiene

$$\begin{aligned} \nabla \beta_{jk} &= -E \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = - \sum_{i=1}^n E \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k} \\ &= \sum_{i=1}^n \frac{x_{ij}x_{ik}}{V[Y_i g'(\mu_i)^2]}, \end{aligned} \quad (8.33)$$

Después de aplicar (8.10) en  $\nabla \beta_{jk}$ , se comprobó que

$$\nabla \beta = X^t W(\beta)^{-1} X,$$

donde  $W = \text{diag } V[Y_i g'(\mu_i)^2]; i = 1, \dots, n$ , si se escribe en notación matricial los términos de la derecha, que aparecen en la ecuación (8.11), se obtiene el vector de puntuaciones. Es decir,

$$U \beta = X^t W^{-1} Z^* \beta,$$

donde

$$\begin{aligned} Z^* \beta_{n \times 1} &= [Y_1 - \mu_1 \ g'(\mu_1)^t; \dots; Y_n - \mu_n \ g'(\mu_n)^t]^t; i = 1, \dots, n \\ &= \text{diag } g'(\mu_1), \dots, g'(\mu_n) \ [Y - \mu]^t. \end{aligned}$$

Sustituyendo en (8.32), el lado izquierdo de la igualdad queda de la siguiente forma

$$\nabla \beta^{r-1} \beta^r = X^t W^{-1} X \beta^r,$$

mientras que el lado derecho se transforma en

$$\begin{aligned}
 & \nabla \beta^{r-1} \beta^{r-1} + U \beta^{r-1} \\
 &= X^t W \beta^{r-1}^{-1} X \beta^{r-1} + X^t W \beta^{r-1}^{-1} Z^*(\beta^{r-1}) \\
 &= X^t W \beta^{r-1}^{-1} X \beta^{r-1} + \text{diag } g' \mu_1, \dots, g' \mu_n - Y - \mu \\
 &= X^t W \beta^{r-1}^{-1} Z(\beta^{r-1})
 \end{aligned}$$

Igualando nuevamente ambos lados, y operando, se obtiene

$$\beta^r = X^t W \beta^{r-1}^{-1} X^{-1} X^t W \beta^{r-1}^{-1} Z \beta^{r-1} \quad (8.34)$$

#### 8.4.4 Algoritmo de estimación

Para estimar los valores del vector de parámetros  $\beta$  de un MLG se emplea el método de máxima verosimilitud. Este método presenta muchas propiedades óptimas, tales como, consistencia y eficiencia asintótica, siendo este más preferido y frecuentemente utilizado computacionalmente. El método de máxima verosimilitud será presentado más detalladamente en a continuación.

El algoritmo de estimación de máxima verosimilitud fue desarrollado por Nelder e Wedderburn (1972) en base a un método semejante al de Newton – Raphson, conocido como el método de score de Fisher. Una principal diferencia en relación al modelo clásico de regresión es que las ecuaciones de máxima verosimilitud son no-lineales.

El método consiste en resolver un sistema  $U \beta = 0$ , en que  $U \beta$  es conocido como función score de  $l \beta$ , entonces

$$U \beta = \frac{\partial l \beta}{\partial \beta},$$

En vez de utilizar la matriz de información de Fisher

$$\mathfrak{I} = -E \frac{\partial^2 l \beta}{\partial \beta_j \partial \beta_k} = -E \frac{\partial U \beta}{\partial \beta}.$$

Expandiendo la función en series de Taylor en términos de primera orden, se tiene:

$$U \beta^{(m+1)} = U \beta^{(m)} + \frac{\partial U \beta^{(m)}}{\partial \beta} \beta^{(m+1)} - \beta^{(m)} = 0$$

o

$$\beta^{(m+1)} = \beta^{(m)} - \left( \frac{\partial U \beta^{(m)}}{\partial \beta} \right)^{-1} U \beta^{(m)},$$

donde el índice  $(m)$  significa el valor del término en la  $m$ -ésima iteración. Este es el método de Newton – Raphson para el cálculo de la estimación de máximo verosímil  $\beta$  de  $\beta$ . El método de Fisher Scoring (1925) se obtiene sustituyendo  $-\frac{\partial U \beta^{(m)}}{\partial \beta}$  por su valor esperado  $\mathfrak{I}$ .

Considere la componente sistemática dada por:

$$\eta_i = g(\mu_i) = \sum_{r=1}^p x_{ir} \beta_r = x_i^T \beta,$$

donde  $x_i^T$  es la  $i$ -ésima fila de  $X$ .

El log – verosímil es dado por

$$l(\beta) = \frac{1}{a(\phi)} \sum_{i=1}^n y_i \theta_i - b(\theta_i) + c(y_i; \phi),$$

derivando  $l(\beta)$  con respecto al vector  $\beta$ , se tiene

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \frac{1}{a(\phi)} \sum_{i=1}^n y_i - b'(\theta_i) \frac{\partial \theta_i}{\partial \beta}.$$

Calculando

$$\frac{\partial \theta_i}{\partial \beta} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta}$$

Por la regla de la cadena y utilizando las ecuaciones (7.7) y (7.8), se obtiene

$$\mu_i = b'(\theta_i) \quad \text{y} \quad V(\mu) = b''(\theta_i) = \frac{\partial \mu_i}{\partial \beta_i},$$

como  $x_i^T$  es la  $i$ -ésima fila de  $X$  y  $\eta_i = x_i^T \beta$ , se tiene:

$$\frac{\partial \eta_i}{\partial \beta} = x_i,$$

donde  $x_i$  es un vector columna  $p \times 1$ , entonces

$$\frac{\partial \mu_i}{\partial \eta_i} = g'(\mu_i)^{-1}.$$

Finalmente la función score es expresada como,

$$U \beta = \frac{\partial l}{\partial \beta} = \frac{1}{a(\phi)} \sum_{i=1}^n y_i - b'(\theta_i) \frac{1}{V(\mu_i)g'(\mu_i)} x_i$$

La matriz de información para  $\beta$  es dada por,

$$\Sigma = X^T W X,$$

donde  $W$  es la matriz diagonal de ponderación definido por

$$w_{ii} = \frac{1}{a(\phi_i)} V_i^{-1} g'(\mu_i)^{-2}.$$

En tanto la función score usando la matriz de ponderación, es dada por

$$U \beta = X^T W z^*,$$

con  $z^*$  es un vector de dimensión  $n \times 1$  expresado por

$$z_i^* = y_i - \mu_i - \frac{\partial g}{\partial \mu_i} \cdot$$

Con éstos resultados, el algoritmo Fisher Scoring para calcular un EMV de  $\beta$  es expresado por

$$\beta^{(m+1)} = \beta^{(m)} + (X^T W^m X)^{-1} X^T W^m z^{*(m)},$$

finalmente, colocando  $(X^T W^m X)^{-1}$  en evidencia se tiene,

$$\beta^{(m+1)} = (X^T W^{-m} X)^{-1} X^T W^{-m} y^{*(m)}, \quad (8.35)$$

en que  $y^{*(m)}$  es la variable respuesta modificada denotada por

$$y^{*(m)} = X\beta^{(m)} + z^{*(m)}.$$

Por tanto la solución de ecuaciones de máxima verosimilitud equivale a calcular repetidamente una regresión lineal ponderada de la variable dependiente modificada  $y^*$  sobre  $X$ , con matriz de ponderación  $W$ . Note que, cuando mayor sea la varianza de observaciones, menor será su ponderación en el cálculo de las estimaciones de los parámetros.

Los programas computacionales de ajuste de MLG usan el método de Fisher Scoring para el cálculo de las estimaciones de los  $\beta$ 's, pues en el método de Newton – Raphson existe una mayor probabilidad de no convergencia del algoritmo.

## ENFOQUE BAYESIANO

### 9. PARADIGMA BAYESIANO

#### 9.1 FUNDAMENTOS DEL ENFOQUE

El marco teórico en que se aplica la inferencia bayesiana es similar a la clásica: hay un parámetro poblacional respecto al cual se desea realizar inferencias y se tiene un modelo que determina la probabilidad de observar diferentes valores de  $\theta$ , bajo diferentes valores de los parámetros. Sin embargo, la diferencia fundamental es que la inferencia bayesiana considera al parámetro como una variable aleatoria. Esto parecería que no tiene demasiada importancia, pero realmente si lo tiene pues conduce a una aproximación diferente para realizar el modelamiento del problema y la inferencia propiamente dicha<sup>7</sup>.

La metodología bayesiana está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el Teorema de Bayes. En esencia, la inferencia bayesiana está basada en la distribución de probabilidad del parámetro dado los datos (distribución posterior de probabilidad  $f(\theta|y)$ , en lugar de la distribución de los datos dado el parámetro. Esta diferencia conduce a inferencias mucho más naturales, lo único que se requiere para el proceso de inferencia bayesiana es la especificación previa de una distribución previa de probabilidad  $f(\theta)$ , la cual representa el conocimiento acerca del parámetro antes de obtener cualquier información respecto a los datos.

---

<sup>7</sup> Citas comunes por sus impulsores contemporáneos “Jeffreys, de Finetti, Good, Savage, Lindley, Zellner”. En el libro Robert C., and Casella G. 2004 “Monte Carlo Statistical methods”. Second Edition. Springer.

## 9.2 DESARROLLO DEL TEOREMA DE BAYES

Sea  $y = y_1, \dots, y_n$  un vector de " $n$ " observaciones cuya distribución de probabilidad  $f(y|\theta)$  depende de  $k$  parámetros involucrados en el vector  $\theta = \theta_1, \dots, \theta_k$ . Supóngase también que tiene una distribución de probabilidades  $f(\theta)$ . Entonces, la distribución conjunta de  $\theta$  e  $y$  es:

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

donde la distribución de probabilidad condicional de  $\theta$  dado el vector de observaciones  $y$  resulta:

$$f(\theta|y) = \frac{f(y|\theta) f(\theta)}{f(y)}$$

con  $f(y) \neq 0$

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

$$f(y) = \begin{cases} \int_{\Theta} f(y|\theta) f(\theta) d\theta & \text{si } \theta \text{ es continuo} \\ \sum_{\theta} f(y|\theta) f(\theta) & \text{si } \theta \text{ es discreto} \end{cases}$$

donde la suma o integral es tomada sobre el espacio paramétrico de  $\theta$ , de este modo, el teorema de Bayes puede ser escrito como:

$$f(\theta|y) = c f(y|\theta) f(\theta) \propto f(y|\theta) f(\theta)$$

donde:

$f(\theta)$  representa lo que es conocido de  $\theta$  antes de recolectar los datos y es llamada la distribución previa de  $\theta$ ;

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

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

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

$$f(\theta|y) \propto l(\theta|y) f(\theta). \quad (9.1)$$

### 9.2.1 Naturaleza secuencial del Teorema de Bayes

El teorema en (9.1) es atractivo debido a que proporciona una formulación matemática de cómo el conocimiento previo puede ser combinado con un nuevo conocimiento.

En efecto, el teorema permite continuamente actualizar la información sobre un conjunto de parámetros  $\theta$  cuando se toman más observaciones. Sea  $y_1$  una muestra inicial, entonces por (9.1) dada anteriormente se tiene:

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

Si se obtiene una 2da muestra de observaciones  $y_2$ , distribuida independientemente de la 1ra muestra de observaciones  $y_1$ , entonces;

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

$$f(\theta | y_1, y_2) \propto f(\theta | y_1) l(\theta | y_2)$$

$$\theta | y_1, y_2 \propto f(\theta | y_1) l(\theta | y_2)$$

$$\propto f(\theta) l(\theta | y_1) l(\theta | y_2)$$

pues:

$$f(y_1, y_2 | \theta) = f(y_1 | \theta) f(y_2 | \theta) \text{ con } y_1, y_2 \text{ independientes.}$$

$$l(\theta | y_1, y_2) = l(\theta | y_1) l(\theta | y_2) \text{ las verosimilitudes también son independientes.}$$

De esta manera, la distribución posterior obtenida con la primera muestra se convierte en la nueva distribución previa para ser corregida por la segunda muestra.

Este proceso puede repetirse indefinidamente. Así, si se tienen "r" muestras independientes, la distribución posterior puede ser re - calculada secuencialmente para cada muestra de la siguiente manera:

$$f(\theta | y_1, y_2, \dots, y_n) \propto l(\theta | y_n) f(\theta | y_1, y_2, \dots, y_{n-1}) \quad \text{para } n = 2, 3, \dots, r$$

Nótese que  $\theta | y_1, y_2, \dots, y_n$  también puede obtenerse partiendo de  $f(\theta)$  y considerando al total de las "r" muestras como una sola gran muestra.

### 9.3 FUNCIÓN DE VEROSIMILITUD

La función de verosimilitud<sup>8</sup>  $l(\theta|y)$  juega un rol muy importante en la fórmula de Bayes, ya que es la función a través de la cual los datos  $y$  modifican el conocimiento previo de  $\theta$ ; puede por tanto ser considerado como representante de la información sobre  $\theta$  que viene en los datos.

La función de verosimilitud es definida con una constante multiplicativa, esto quiere decir que la multiplicación de la función de verosimilitud por una constante deja a la verosimilitud sin cambio, puesto que multiplicar la función de verosimilitud por una constante arbitraria no tendrá efecto en la distribución posterior de  $\theta$ . La constante se cancelará cuando se normalice el producto en el lado derecho de  $f(\theta|y) = c f(y|\theta) f(\theta)$ .

Por último es conveniente señalar que la información muestral  $y$  por lo general será introducida en el modelo a través de estadísticas suficientes para  $\theta$ , dado que estas contienen toda la información referente a los datos. Así, dado un conjunto de estadísticas suficientes  $t(y)$  para los parámetros en  $\theta$ ,  $f(y|\theta)$  puede ser intercambiada por  $f(t(y)|\theta)$ , para lo cual bastara con calcular la distribución condicional de  $y$  dado  $\theta$ .

#### **Ejemplo 8: Función de verosimilitud para un MLG Poisson**

El modelo lineal generalizado Poisson asume que  $y$  se distribuye Poisson con media  $\mu$  y con función enlace logarítmico, tal que  $\ln \mu = X\beta$ . La distribución muestral para los datos  $y = y_1, \dots, y_n$  es

---

<sup>8</sup> La función de Verosimilitud en el enfoque Bayesiano tiene un tinte distinto que en el enfoque Clásico.

$$f(y|\beta) = \prod_{i=1}^n \frac{1}{y_i!} e^{\exp(\eta_i)} \exp(\eta_i)^{y_i}$$

donde  $\eta_i = X\beta_i$  es el predictor lineal para el  $i$ -ésimo caso. Cuando se considera la distribución posterior, el factor  $\frac{1}{y_i!}$  es inducido a una constante arbitraria.

## 9.4 DISTRIBUCION PREVIA CONJUGADA

Tanto  $f(\theta)$  como  $f(\theta|y)$  son distribuciones de probabilidad sobre  $\theta$ , la primera sólo incorpora información previa y la segunda actualiza dicha información con la información muestral que se pueda obtener. Si bien se mencionó que la elección de una u otra distribución de probabilidad para modelar la incertidumbre sobre  $\theta$  no resulta crucial en tanto sea factible eliciar con cualquiera de ellas una distribución previa, resulta conveniente tanto para el análisis como desde un punto de vista computacional el que  $f(\theta)$  y  $f(\theta|y)$  pertenezcan a la misma familia.

### 9.4.1 Definición

Sea  $\mathcal{P} = \{f(\theta) : \theta \in \Theta\}$  una familia paramétrica. Una clase  $P$  de distribuciones de probabilidad  $\mathcal{F}$  es una familia conjugada para  $\mathcal{P}$  si para todo  $f(\theta) \in \mathcal{P}$  y  $f(\theta) \in \mathcal{F}$  se cumple  $f(\theta|y) \in \mathcal{F}$ .

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

#### 9.4.2 Distribución previa conjugada, Familia exponencial y Suficiencia

Para relacionar familia conjugada con los clásicos conceptos de familia exponencial y suficiencia vistos en el capítulo 7, las distribuciones que subyacen de una familia exponencial, tienen una distribución previa conjugada.

Sea la variable aleatoria  $Y$  con función de densidad o función de masa de probabilidad  $f(y|\theta, \phi)$ , la cual depende de los parámetros  $\theta$  y  $\phi$ . Esta función, o más específicamente esta familia de funciones de densidad o de masa de probabilidad, pertenecen a la familia exponencial si puede ser escrita en la forma

$$f(y|\theta, \phi) = \exp \frac{(y\theta - b(\theta))}{a(\phi)} + c(y; \phi)$$

Cuando una f.d.p o f.m.p no tiene un parámetro de dispersión entonces  $a(\phi) = 1$ . En este caso, la expresión (9.1) se reduce a

$$f(y|\theta) = e^{y\theta - b(\theta) + c(y)},$$

la que también puede ser escrita como

$$f(y|\theta) = e^{y\theta} e^{-b(\theta)} e^{c(y)}$$

Alternativamente

$$f(y|\theta) = a(\theta) b(y) e^{y\theta}$$

donde  $a(\theta) = e^{-b(\theta)}$ ,  $b(y) = e^{c(y)}$ ,  $a(\theta) > 0$  y  $b(y) > 0$  para todo  $\theta$  e  $y$ , respectivamente.

La verosimilitud correspondiente para una muestra aleatoria  $y_1, \dots, y_n$  independiente e idénticamente distribuida, es

$$f(y|\theta) = \prod_{i=1}^n b(y_i|\theta) = a(\theta)^n e^{\theta} \prod_{i=1}^n y_i,$$

como función de  $\theta$

$$f(y|\theta) \propto a(\theta)^n e^{\theta t(y)}, \quad (9.2)$$

donde  $\sum_{i=1}^n y_i = t(y)$  es la estadística suficiente para el parámetro, porque la verosimilitud para  $\theta$  depende sobre los datos  $y$  sólo a través de  $t(y)$ .

Las estadísticas suficientes son manipuladas fuertemente en las operaciones algebraicas de verosimilitud y distribuciones posteriores. Si la distribución previa es especificada como:

$$f(\theta) \propto a(\theta)^\eta e^{\theta v}.$$

Entonces, la distribución posterior es

$$f(\theta|y) \propto a(\theta)^{\eta+n} e^{\theta(t(y)+v)} \quad (9.3)$$

Lo cual demuestra que la distribución previa es conjugada. En general la familia exponencial representa una clase especial de distribuciones que tienen una distribución previa conjugada natural<sup>9</sup>.

---

<sup>9</sup> Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin. 1997 "Bayesian Data Analysis". Second Edition. Chapman and Hall.

### **Ejemplo 9: Previa conjugada para el modelo Poisson<sup>10</sup>**

Sea  $\mu$  la tasa de ocurrencia o el número esperado de veces que un evento ocurre en un periodo de tiempo dado. También se define la variable aleatoria  $y$  como el número de veces que un evento ocurre. La relación entre la frecuencia esperada,  $\mu$ , y la probabilidad de observar la frecuencia  $y$  es especificada por la distribución Poisson como sigue

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \quad y = 0, 1, 2, \dots$$

La verosimilitud correspondiente para una muestra aleatoria  $y_1, \dots, y_n$  independiente e idénticamente distribuida, de una distribución Poisson ( $\mu$ ), es el producto de las verosimilitudes originales:

$$\begin{aligned} f(y_1, \dots, y_n|\mu) &= \prod_{i=1}^n f(y_i|\mu) \\ &= \prod_{i=1}^n \frac{e^{-\mu}\mu^{y_i}}{y_i!} \\ &= \prod_{i=1}^n y_i! e^{-ny}\mu^{\sum_{i=1}^n y_i} \end{aligned}$$

$$f(y|\mu) \propto \mu^{t(y)} e^{-ny},$$

donde  $\sum_{i=1}^n y_i = t(y)$  es la estadística suficiente. Si lo anterior se escribe en términos de la familia exponencial,

---

<sup>10</sup> El ejemplo también apoya el desarrollo de la relación de la distribución Poisson con la distribución Binomial Negativa en el enfoque Clásico

$$f(y|\mu) \propto e^{-\mu} \mu^y$$

donde el parámetro natural es  $\theta = \ln(\mu)$ , y la distribución previa conjugada natural es

$$f(\mu) \propto e^{-\eta} \mu^{\eta}$$

Con los hiperparámetros indexados  $(\eta, v)$ , para poner el argumento de otra manera, la verosimilitud es de la forma  $\mu^a e^{-b\theta}$ , y la distribución previa conjugada debe ser de la forma

$$f(\mu) \propto e^{-\beta\mu} \mu^{\alpha-1} \quad (9.4)$$

La cual es una distribución gamma con parámetros  $\alpha$  y  $\beta$ . Comparando  $f(\mu)$  y  $f(y|\mu)$  se revela que la densidad previa es, en algún sentido equivalente para un conteo total de  $\alpha - 1$  en  $\beta$  previas observaciones.

Con esta distribución previa conjugada, la distribución posterior es

$$\frac{\mu}{y} \sim \text{Gamma } \alpha + ny, \beta + n .$$

De la ecuación de bayes, y con una sola observación para  $y$ , se tiene

$$f(y) = \frac{\text{Poisson } y | \mu \times \text{Gamma } \mu | \alpha, \beta}{\text{Gamma } \mu | \alpha + y, 1 + \beta}$$

$$f(y) = \frac{\frac{e^{-\mu} \mu^y}{y!} \frac{\beta^\alpha}{\Gamma \alpha} e^{-\beta\mu} \mu^{\alpha-1}}{\frac{1 + \beta^{\alpha+y}}{\Gamma \alpha + y} e^{-1 + \beta^\mu} \mu^{\alpha+y-1}}$$

Realizando operaciones algebraicas

$$f(y) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)y!} \frac{\beta^\alpha}{1 + \beta^{\alpha+y}},$$

con  $\Gamma(\alpha + y) = \alpha + y - 1$ , se obtiene

$$f(y) = \frac{\alpha + y - 1}{y} \left( \frac{\beta}{1 + \beta} \right)^\alpha \left( \frac{1}{1 + \beta} \right)^y,$$

que tiene forma de una distribución Binomial Negativa  $y \sim BN(\alpha, \beta)$ .

Para otras distribuciones y sus respectivas previas conjugadas ver Apendice A de Bayesian Data Analysis, Gelman, Carlin Stern y Rubin (1992) Segunda Edición.

## 9.5 DISTRIBUCIONES PREVIAS NO INFORMATIVAS

Se habla frecuentemente de una distribución previa no informativa cuando esta es localmente plana, vaga y difusa con respecto a la función de verosimilitud. En algunos casos las previas no informativas llevan a distribuciones posteriores impropias (densidad posterior no integrable), lo que lleva a hacer inferencias con distribuciones a posteriores impropias. En adición, una previa no informativa a menudo es no invariantes bajo transformación; esto es, una previa puede ser no informativa en una parametrización pero no necesariamente no informativa si una transformación es aplicada<sup>11</sup>. Una común selección para una previa no informativa es la previa plana, que es una distribución previa que asigna igual verosimilitud sobre todos los valores posibles del parámetro  $\theta$ .

---

<sup>11</sup> Bernardo, J. & Smith, A. 1994. "Bayesian Theory". First Edition. Wiley & Sons, New York.

### 9.5.1 Principio de Invarianza de Jeffreys

Jeffreys (1961) definió una distribución previa no informativa considerando transformaciones uno a uno del parámetro:  $\phi = h(\theta)$ , por transformación de variables la densidad previa  $f(\theta)$  es equivalente a la densidad previa para  $\phi$ :

$$f(\phi) = f(\theta) = h^{-1}(\phi) \frac{d\theta}{d\phi}$$

$$f(\phi) = f(\theta) \frac{d\theta}{d\phi}$$

Jeffreys propuso como previa no informativa a:

$$f(\theta) \propto \sqrt{J(\theta)} \quad (9.5)$$

Donde  $J(\theta)$  es la matriz de información de Fisher

$$J(\theta) = -E \frac{\partial^2 \log f(\theta)}{\partial \theta^2}$$

*Teorema.* La distribución previa no informativa de Jeffreys  $f(\theta) \propto J(\theta)^{1/2}$  es invariantante ante transformaciones uno-a-uno, esto es, si  $\phi = h(\theta)$  es una transformación uno-a-uno de  $\theta$  entonces la distribución previa de  $\phi$  es  $f(\phi) \propto J(\phi)^{1/2}$ .

*Demostración.* Para verificar que la previa de Jeffreys es invariantante bajo transformaciones, se evalúa  $J(\phi)$  para  $\phi = h(\theta)$ , la cual es una transformación uno a uno de  $\theta$ , entonces:

$$\frac{\partial \log f(y|\theta)}{\partial \phi} = \frac{\partial \log f(y|h(\theta))}{\partial \theta} \frac{d\theta}{d\phi}$$

donde  $\theta = h^{-1}(\phi)$  es la inversa de la transformación uno a uno de  $\theta$ , Para obtener la información de Fisher de  $\phi$  se calcula:

$$\frac{\partial^2 \log f(y|\phi)}{\partial \phi^2} = \frac{\partial \log f(y|h^{-1}(\phi))}{\partial \theta} \frac{\partial^2 \theta}{\partial \phi^2} + \frac{\partial^2 \log f(y|h^{-1}(\phi))}{\partial \theta^2} \frac{d\theta}{d\phi}^2$$

multiplicando ambos miembros por -1 y calculando esperanza respecto a  $f(y|\theta)$

$$J(\phi) = -E \left[ \frac{\partial \log f(y|\phi)}{\partial \phi} \right]^2 = J(\theta) \left( \frac{d\theta}{d\phi} \right)^2$$

pero se tiene que

$$E \left[ \frac{\partial \log f(y|\theta)}{\partial \theta} \right]$$

$$= E \left[ \frac{\frac{\partial}{\partial \theta} f(y|\theta)}{f(y|\theta)} \right]$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial \theta} f(y|\theta)}{f(y|\theta)} f(y|\theta) dy$$

$$= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(y|\theta) dy$$

$\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(y|\theta) dy$  por las condiciones de

regularidad)

$$\frac{\partial}{\partial \theta} 1 = 0$$

$$J(\phi) = J(\theta) \left| \frac{d\theta}{d\phi} \right|^2$$

entonces,

$$J(\phi)^{1/2} = J(\theta)^{1/2} \left| \frac{d\theta}{d\phi} \right|$$

pero  $\left| \frac{d\theta}{d\phi} \right|$  es el valor absoluto del jacobiano de la transformación inversa por lo que si  $f(\theta) \propto J(\theta)^{1/2}$  entonces

$$f(\phi) \propto \overline{J(\theta(\phi))} \left| \frac{d\theta}{d\phi} \right| = J(\phi)^{1/2} \quad (9.6)$$

y por lo tanto la distribución a priori de Jeffreys es invariantante ante transformaciones uno-a-uno.

#### **Ejemplo 10: Previa no Informativa de Jeffreys para un modelo Poisson**

Sea la variable aleatoria  $y$  con una distribución Poisson ( $\mu$ )

$$f(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

$$\log f(y|\mu) = \log \frac{1}{y!} + y \log \mu - \mu$$

$$\frac{d \log f(y|\mu)}{d\mu} = \frac{y}{\mu}$$

$$\frac{d^2 \log f(y|\mu)}{d\mu^2} = -\frac{y}{\mu^2}$$

$$-E\left[-\frac{y}{\mu^2}\right] = \frac{1}{\mu^2} E(y)$$

$$J(\mu) = \frac{1}{\mu}$$

$$f(\mu) \propto \overline{J(\phi)} = \frac{1}{\mu} \quad (9.7)$$

## 9.6 DISTRIBUCION PREDICTIVA

El carácter de predictibilidad que otorga el método bayesiano es destacable y básicamente consiste en reconocer que el proceso de razonamiento es idéntico siempre. El conocimiento previo se actualiza con información muestral para producir nuevas creencias sobre el sistema. Tal y como se observó en el enunciado del Teorema de Bayes, el factor de proporcionalidad que convierte en igualdad el ajuste del juicio posterior mediante la verosimilitud y la previa es la distribución marginal

$$f(y) = \begin{cases} \int_{\Theta} f(y|\theta) f(\theta) d\theta & \text{si } \theta \text{ es continuo} \\ f(y|\theta) f(\theta) & \text{si } \theta \text{ es discreto} \end{cases}$$

A  $f(y)$  se denomina distribución predictiva previa (o inicial), y describe el conocimiento acerca de una observación futura  $y$  basado únicamente en la información contenida en  $f(\theta)$ . Nótese que  $f(y)$  no depende ya de  $\theta$ .

Es previa porque no es condicional sobre las observaciones, y predictiva porque es la distribución para una cantidad que es observada.

Después que los datos son observados, se puede predecir una observación desconocida  $y$  la cual es la distribución predictiva posterior, posterior porque es condicional sobre las observaciones  $y$  y predictiva porque es una predicción para una observación  $y$

$$\begin{aligned} f(y|y) &= \int_{\Theta} f(y|\theta, y)d\theta \\ &= \int_{\Theta} f(y|\theta, y)f(\theta|y)d\theta \\ &= \int_{\Theta} f(y|\theta)f(\theta)d\theta \end{aligned}$$

La segunda y tercera línea exhiben la distribución posterior predictiva como un promedio de la predictiva condicional sobre la distribución posterior de  $\theta$ .

La última línea concluye de esa manera pues  $y$  e  $y$  son condicionalmente independientes dado  $\theta$  en el modelo.

## 9.7 DISTRIBUCION POSTERIOR

Dado que la distribución posterior, contiene toda la información concerniente al parámetro de interés  $\theta$  (información a priori y muestral), cualquier inferencia con respecto a  $\theta$  consistirá en afirmaciones hechas a partir de dicha distribución.

Es el resultado final del análisis bayesiano, toda la información se obtiene de ella; estimaciones puntuales: media, mediana, moda...Intervalos de Credibilidad.

Existen dificultades prácticas para obtener la distribución posterior, sobre todo en modelos complejos. Más adelante se presenta una descripción para subsanar tales dificultades.

### **Ejemplo 12: Distribución posterior para un modelo Poisson**

Continuando el ejemplo 11,

$$f(y/\mu) \propto \mu^{t(y)} e^{-n\mu}$$

$$f(\mu) \propto \overline{J(\phi)} = \frac{1}{\mu}$$

Claramente se ve que la distribución previa no es propia, sin embargo haciendo el cálculo de la distribución posterior  $f(\mu|y) \propto f(y|\mu) f(\mu)$ , se ve que

$$f(\mu|y) \propto \mu^{t(y)} e^{-n\mu} \frac{1}{\mu}$$

$$f(\mu|y) \propto e^{-n\mu} \mu^{\sum_{i=1}^n y_i - 1/2} \quad (9.8)$$

La cual es el kernel de una distribución Gamma con parámetros  $(\sum_{i=1}^n y_i + 1/2, n)$ .

### **9.8 METODOS MCMC**

Como se revisó hasta ahora en este capítulo, la inferencia Bayesiana está basada en el análisis de la distribución posterior porque esta contiene toda la información sobre el parámetro a ser estimado condicional a los datos observados  $y$ . En general en inferencia Bayesiana es útil resumir la información que está contenida en la distribución posterior, y estos resúmenes típicamente toman la forma de esperanzas de funciones particulares de los parámetros, esto es,

$$I = E[g(\theta)] = \int g(\theta) f(\theta|y) d\theta$$

mas, el problema general de inferencia bayesiana consiste en calcular estos valores esperados según la distribución posterior de  $\theta$ . Usualmente es muy complicado o imposible evaluar  $I$  en forma analítica, incluso las técnicas numéricas de cuadratura u otras para aproximar podrán presentar problemas, más aun si el parámetro es multidimensional. Por este motivo se hace necesario utilizar métodos aproximados para obtener estas integrales.

En estos últimos años la inferencia bayesiana ha experimentado un gran avance debido a la introducción de técnicas de simulación que permiten en forma relativamente simple obtener una muestra de la distribución objetivo<sup>12</sup>. En particular, los métodos conocidos como Cadenas de Markov via Monte Carlo (MCMC) son actualmente muy utilizados. La racionalidad de estos métodos subyace en diseñar iterativamente una cadena de Markov para  $\theta$  de tal manera que  $f(\theta)$  sea su distribución ergódica estacionaria. Empezando en algún estado inicial  $\theta^0$  la idea es simular un número suficientemente grande  $M$  de transiciones bajo la cadena de Markov y registrar los correspondientes estados simulados  $\theta_j$ . Luego, bajo ciertas condiciones de regularidad, es posible mostrar que la media muestral ergódica

$$I = \frac{1}{M} \sum_{j=1}^M g(\theta_j)$$

converge a la integral  $I$  deseada. En otras palabras,  $I$  nos provee de una buena aproximación para  $I$ . El reto de los métodos MCMC consiste entonces en precisar una cadena de Markov adecuada con la distribución posterior  $f(\theta)$  y como su distribución estacionaria o ergódica y decidir cuándo detener la simulación.

---

<sup>12</sup> También denominada distribución posterior.

Existe un teorema de ergodicidad que es el equivalente a la ley fuerte de los grandes números pero con cadenas de Markov. Establece que cualquier función de la distribución se puede estimar mediante una muestra de la cadena de Markov obtenida a partir de un estado ergódico. Así, tras alcanzar el estado ergódico se obtiene una muestra válida de la distribución posterior y se emplea los datos para estimar medias, varianzas o cuantiles.

En Estadística Bayesiana el problema es el inverso: cómo construir la cadena de Markov para obtener una muestra de una distribución dada. Existen dos técnicas básicas<sup>13</sup>: El muestreador de Metropolis-Hastings y el muestreador de Gibbs.

Describamos ahora uno de los métodos MCMC más populares conocido como el algoritmo de Metrópolis-Hastings.

### 9.8.1 ALGORITMO DE METRÓPOLIS-HASTING

Metrópolis - Hasting es un método Monte Carlo vía cadenas de Markov, en su forma más simple lo publica Metrópolis (1953) y lo describe de la siguiente manera:

Metrópolis es un algoritmo simple, práctico y puede ser usado para obtener muestras aleatorias desde cualquier distribución propuesta. Suponer que se quiere obtener  $T$  muestras desde una distribución univariante con función de densidad de probabilidad  $f(\theta|y)$ . Suponer que  $\theta^t$  es la  $t$ -ésima muestra de  $f$ . Para usar el algoritmo de Metrópolis, se necesita un valor inicial  $\theta^0$  y una densidad simétrica propuesta (o distribución generadora de candidatos)  $q(\theta^{t+1}|\theta^t)$ . Para la  $t$ -ésima iteración, el algoritmo genera una muestra desde  $q(\cdot|\cdot)$  basada sobre la muestra actual  $\theta^t$ , y esta toma la decisión de aceptar o

---

<sup>13</sup> Existen otras técnicas que se aplican según la estructura de la información con la que se trabaja.

rechazar la nueva muestra. Si la nueva muestra es aceptada, el algoritmo se repite generando una nueva muestra. Si la nueva muestra es rechazada, el algoritmo vuelve a empezar en su punto actual y se repite hasta que la muestra sea aceptada. El algoritmo puede ser repetido tanta veces se requiera, en la práctica, se tiene que decidir el número total de muestras que se necesita en adelante y parar de muestrear después de que muchas iteraciones hayan sido completadas.

Suponer  $q(\theta_{nuevo} | \theta^t)$  es una distribución simétrica. La distribución propuesta debiera ser una distribución fácil de muestrear, y debe ser tal que  $q(\theta_{nuevo} | \theta^t) = q(\theta^t | \theta_{nuevo})$ , significa que la verosimilitud de incremento para  $\theta_{nuevo}$  desde  $\theta^t$  es la misma como la verosimilitud de incremento para  $\theta^t$  de  $\theta_{nuevo}$ . La más común selección de distribución propuesta es la distribución normal  $N(\theta^t, \sigma)$  con  $\sigma$  fijo. El algoritmo de metrópolis puede ser resumido en los siguientes pasos:

1. Establecer  $t = 0$ . Elegir un punto inicial  $\theta^0$  este puede ser un punto arbitrario mientras se cumpla  $f(\theta^0) > 0$ .
2. Generar una nueva muestra,  $\theta_{nuevo}$ , usando la distribución propuesta  $q(\cdot | \theta^t)$ .
3. Calcular la siguiente cantidad

$$r = \min \frac{f(\theta_{nuevo} | y)}{f(\theta^t | y)}, 1$$

4. Muestrear  $\mu$  desde la distribución uniforme  $U(0, 1)$ .
5. Generar  $\theta^{t+1} = \theta_{nuevo}$  si  $\mu < r$ ; en otro caso  $\theta^{t+1} = \theta^t$ .

6. Establecer  $t = t + 1$ . Si  $t < T$ , el número de muestras deseadas, retornar al paso 2. En otro caso, parar.

Nótese que el número de iteraciones se mantiene creciente, a pesar de que la muestra propuesta es aceptada o no.

El algoritmo define una cadena de variables aleatorias cuya distribución convergerá hacia la distribución deseada  $f(\theta|y)$ , y así desde algunos puntos adelante, la cadena de muestras; es una muestra desde la distribución de interés. En terminología de cadena de Markov; ésta distribución es llamada la distribución estacionaria de la cadena, y en estadística Bayesiana, esta es la distribución posterior de los parámetros. La razón del trabajo del algoritmo de Metrópolis va mas allá del ámbito que cubre ésta tesis, pero se pude encontrar más descripciones a detalle con pruebas en textos generales como Roberts (1996) y Lui (2001).

Generalmente no se dispone de una distribución simétrica propuesta  $q(\theta^{t+1}|\theta^t)$ , a tal situación se recurre a desarrollo del algoritmo de Metrópolis Hasting, que fue propuesto por Hasting (1970), el cual propone una distribución asimétrica:  $q(\theta_{nuevo}|\theta^t) \neq q(\theta^t|\theta_{nuevo})$ . La diferencia en esta implementación viene a calcular la razón de densidades:

$$r = \min \frac{f(\theta_{nuevo}|y) q(\theta^t|\theta_{nuevo})}{f(\theta^t|y) q(\theta_{nuevo}|\theta^t)}, 1$$

Los otros pasos permanecen igual.

La extensión del algoritmo de Metrópolis con dimensión superior  $\theta$  es sencilla. Suponer  $\theta = \theta_1, \theta_2, \dots, \theta_k$  ' es un vector de parámetros. Para empezar el algoritmo de Metrópolis, seleccionamos un valor inicial para cada  $\theta_k$  y usar la

versión multivariante de la distribución propuesta  $q(\cdot | \cdot)$ , tal como una distribución Normal Multivariada, para seleccionar un nuevo parámetro  $k$  – dimensional. Los pasos siguientes permanecen iguales a los descritos previamente, y esta cadena de Markov eventualmente converge hacia la distribución objetivo de  $f(\theta|y)$ .

Utilizando el método de simulación de Metrópolis Hastings se calcula el  $E(\theta|y)$  y  $V(\theta|y)$ .

Para construir la cadena  $\{\theta^{(t)}\}$ , las probabilidades de transición  $p(\theta^{(t+1)}|q|\theta^{(t)})$  vendrán dadas por una distribución arbitraria, (distribución generadora de candidatos),  $q(\theta, \theta')$  tal que  $q(\theta, \theta')d\theta' = 1$ , dados el valor actual  $\theta$ , y el valor candidato  $\theta'$ .

### *Teorema*

La cadena de Markov  $Z$  construida por el algoritmo de Metrópolis – Hasting converge a su distribución ergódica o estacionaria  $\pi$ .

### *Demostración:*

La cadena formada por Metrópolis – Hasting es una cadena de Markov con matriz de transición

$$p_{ij} = \begin{cases} q_{ij}\alpha(i,j) & \text{si } i \neq j \\ 1 - \sum_{k \neq i} q_{ik}\alpha(i,k) & \text{si } i = j \end{cases}$$

Ahora se demostrará que la cadena es reversible. Por ser  $\pi$  la función de probabilidad de la variable aleatoria  $X$ , entonces

a)  $0 \leq \pi_i \leq 1,$

b)  $\sum_i \pi_i = 1.$

Para ver la última propiedad

c)  $\pi_i q_{ij} = \pi_j q_{ji}$

suponga que

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} < 1.$$

Entonces,

$$\pi_i p_{ij} = (\pi_i q_{ij}) \alpha_{i,j}$$

$$= \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}}$$

$$= \pi_j q_{ji}$$

$$= \pi_j q_{ji} \min \left( \frac{\pi_i q_{ij}}{\pi_j q_{ji}}, 1 \right)$$

$$= \pi_j q_{ji}.$$

Si

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} > 1, \text{ entonces } \frac{\pi_i q_{ij}}{\pi_j q_{ji}} < 1,$$

por un razonamiento idéntico al anterior, se afirma que  $\pi_j q_{ji} = \pi_i q_{ij}$ .

Por último, el caso en que

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} = 1$$

Es directo. Con esto se concluye que la cadena es positivo - recurrente y reversible con distribución estacionaria  $\pi$ .

Como  $Z$  es una cadena irreducible, entonces para cualquier pareja de estados  $i, j$

$$\pi_i q_{ij} > 0,$$

y por lo tanto

$$\min 1, \frac{\pi_i q_{ij}}{\pi_j q_{ji}} > 0.$$

Con esto se tiene que la cadena que construye el algoritmo de Metrópolis – Hasting es reversible en el tiempo, aperiódica y recurrente positiva con distribución ergódica o estacionaria  $\pi$ , por lo que el teorema asegurará la convergencia de la cadena hacia su distribución estacionaria.

### 9.8.2 MUESTREADOR DE GIBBS

El muestreador de Gibbs llamado así por Geman y Geman (1984), es un caso especial del algoritmo de Metrópolis – Hasting, donde la distribución propuesta coincide exactamente con la distribución posterior condicional. El muestreador

de Gibbs requiere descomponer la distribución posterior conjunta hacia una distribución condicional completa para cada parámetro en el modelo y entonces muestrea desde ellos. El muestreo puede ser eficiente cuando los parámetros no dependen sobre otro y la distribución condicional completa es fácil de muestrear de ella. Algunos investigadores están a favor de este algoritmo porque este no requiere como instrumento una distribución propuesta como el método de Metrópolis los hace. Sin embargo, mientras se deriva la distribución condicional puede ser relativamente fácil, esto no siempre es posible para encontrar un eficiente camino para muestrear desde estas distribuciones condicionales.

Suponer  $\theta = \theta_1, \theta_2, \dots, \theta_k$  el vector de parámetros,  $f$  y  $\theta$  es la verosimilitud, y  $f(\theta)$  es la distribución previa. La distribución posterior condicional completa de  $f(\theta_i | \theta_j, i \neq j, y)$  es proporcional a la densidad posterior conjunta; esto es:

$$f(\theta_i | \theta_j, i \neq j, y) \propto f \text{ y } \theta \ f(\theta)$$

Por instancia la distribución condicional unidimensional de  $\theta_1$  dado  $\theta_j = \theta_j^*, 2 \leq j \leq k$ , es computado de la siguiente manera:

$$f \theta_i | \theta_j = \theta_j^*, 2 \leq j \leq k, y = f(y | (\theta = \theta_1, \theta_2^*, \dots, \theta_k^*))$$

El muestreador de Gibbs trabaja como sigue:

- 1- Definir  $t = 0$ , y elegir un valor inicial arbitrariamente de  $\theta^0 = \theta_1^0, \dots, \theta_k^0$ .
- 2- Generar cada componente de  $\theta$  como sigue:

- Mostrar  $\theta_1^{t+1}$  desde  $f(\theta_1 | \theta_2^{(t)}, \dots, \theta_k^{(t)}, y)$
  - Mostrar  $\theta_2^{t+1}$  desde  $f(\theta_2 | \theta_1^{t+1}, \theta_3^{(t)}, \dots, \theta_k^{(t)}, y)$
  - .....
  - Mostrar  $\theta_k^{t+1}$  desde  $f(\theta_k | \theta_1^{t+1}, \dots, \theta_{k-1}^{t+1}, y)$
- 3- Generar  $t = t + 1$ . Si  $t < T$ , el número de muestras deseadas, retornar al paso 2. En otro caso, parar.

Para realizar la estimación se utiliza la metodología del Muestreador de Gibbs, algoritmo que extrae muestras en forma sucesiva de las probabilidades condicionales de los parámetros del modelo.

Son ciertas condiciones de regularidad que dicen que la distribución límite de  $\theta^j$  tiende a  $\pi(\theta)$ <sup>14</sup>.

## 9.9 MONITOREO DE CONVERGENCIA

Hasta ahora no existen resultados teóricos de fácil utilización en aplicaciones concretas sobre cuándo se puede considerar estacionaria la cadena de Markov. En especial, sigue siendo un tema de investigación atractivo la búsqueda de métodos que sugieran automáticamente el número  $M$  de etapas iniciales de la cadena que deben ser desechadas, e incluso si es más eficiente utilizar una o varias cadenas independientes para monitorizar esa convergencia; ver, por ejemplo, Gelman y Rubin (1992), Smith y Roberts (1993) y referencias allí citadas.

---

<sup>14</sup> Teorema demostrado anteriormente en éste capítulo.

Uno de los abordajes más comunes es la inspección gráfica de convergencia la cadena, donde el analista observa la trayectoria de una o más cadenas en distintos tiempos, y se afirmará que la convergencia fue alcanzada cuando la o las cadenas monitoreadas permanecen en torno de un mismo punto. Otros criterios más formales también pueden ser utilizados como los métodos propuestos por Gelman y Rubin vistos a continuación.

### 9.9.1 Diagnóstico de convergencia de Gelman y Rubin

En 1992, Gelman y Rubin propusieron una aproximación cuantitativa para monitorear la convergencia de una MCMC; el método se basa en monitorear por separado la convergencia de todas las cantidades de interés de una distribución y cuando la varianza entre las diferentes cadenas no es más grande que la varianza dentro de cada cadena individual, se declara convergencia.

#### 9.9.1.1 DEFINICIÓN

Sea  $\theta = (\theta_1, \dots, \theta_p)$  un vector de parámetros con cierta distribución posterior, supóngase que se está interesado en cada uno de los componentes  $\theta_i$ ,  $i = 1, \dots, p$ . El monitoreo de la convergencia se realiza por separado para cada  $\theta_i$ ; sin pérdida de generalidad, sea  $\theta$  cualquier  $\theta_i$  y supóngase que se tienen  $m$  cadenas paralelas, cada una con longitud  $n$  con  $(\theta_{ij})$ ,  $j = 1, \dots, n$ ;  $i = 1, \dots, m$  y se obtiene la varianza  $B$ , entre las cadenas y la varianza  $W$ , dentro de las cadenas como sigue:

$$B = \frac{n}{m-1} \sum_{i=1}^m \theta_{i..} - \theta_{...}^2,$$

donde  $\theta_{i..} = \frac{1}{n} \sum_{j=1}^n \theta_{ij}$ ,  $\theta_{...} = \frac{1}{m} \sum_{i=1}^m \theta_{i..}$  y

$$W = \frac{1}{m(n-1)} \sum_{i=1}^m \sum_{j=1}^n (\theta_{ij} - \bar{\theta}_{i.})^2.$$

La varianza  $V(\theta|y)$  de la marginal de  $\theta$  que se estima como el promedio ponderado de  $W$  y  $B$ ,

$$V(\theta|y) = \frac{n-1}{n} W + \frac{1}{n} B,$$

Es un estimador insesgado bajo estacionaridad, pero que sobreestima la  $V(\theta|y)$  asumiendo que la distribución previa está apropiadamente sobredispersada. Así, si se tiene una sobredispersión, entonces  $V(\theta|y)$  es un estimador conservador de la varianza de  $\theta$ , ya que cada cadena tiene menos variabilidad que la distribución de todas las cadenas. En el límite, cuando  $n \rightarrow \infty$ ,  $V(\theta|y)$  y  $W$  aproximan a  $V(\theta|y)$  en direcciones opuestas.

Se puede monitorear la convergencia de la cadena de Markov, estimando  $R$ , el factor por el cual la escala de la distribución actual de  $\theta$  puede ser reducido si las simulaciones son continuadas indefinidamente. Esta reducción potencial de escala puede ser estimada usando las anteriores formulas de la siguiente manera:

$$R = \frac{\overline{V(\theta)}}{W} \quad (9.9)$$

y es denominado “estimador de reducción potencial de escala”.

Cuando la cadena converge,  $R$  se acerca a 1, significando que las cadenas de Markov están esencialmente sobreestimadas y provienen de la misma

distribución. Si  $R$  es mayor que 1, entonces se tendrá razón para creer que procediendo con más simulaciones se pueden mejorar las inferencias sobre  $\theta$ .

Se recomienda calcular el “estimador de reducción potencial de escala” para todas las cantidades de interés, descartando la primera mitad de las simulaciones de cada cadena. Si  $R$  no está cercano a 1 para todos los componentes de interés, una buena idea es continuar la simulación. En la práctica se corren simulaciones hasta que los valores de  $R$  son todos menores que 1,1 o 1,2.

Una vez que  $R$  esté cercano a 1 para todas las cantidades de interés, las inferencias se realizan con la muestra obtenida al mezclar las simulaciones de las segundas mitades de las  $m$  cadenas paralelas. Las estimaciones son confiables si están basadas en cadenas múltiples con puntos de inicio sobredispersados. Las estimaciones obtenidas antes de la convergencia serán conservadoras y una vez que la convergencia ha sido alcanzada, éstas serán más precisas. (Gelman 2004).

Cuando se alcanza la convergencia, el numerador y el denominador deberían coincidir, por lo que el estadístico se aproxima a 1.

## 9.10 COMPARACION DE LOS MODELOS

Se puede considerar diferentes modelos para un conjunto de datos, en este acápite se revisa diferentes criterios para la comparación.

Existen una serie de metodologías para compararlos, entre los principales criterios en la inferencia bayesiana se tiene: (deviance information criterion) (DIC) propuesto por Spiegelhalter et al. (2002), el esperado del criterio de información de Akaike (EAIC) y el esperado del criterio de información de

Schwarz o Bayesiano (EBIC)<sup>15</sup>, los dos últimos propuestos en Carlin and Louis (2000). Estos criterios son basados en media posterior del desvío  $E[D(a, b, \lambda, \theta)]$ ; donde

$$E[D(a, b, \lambda, \theta)] = -2 \ln p(y | a, b, \lambda, \theta) = -2 \sum_{i=1}^n \ln P(Y_{ij} = y_{ij} | a, b, \lambda, \theta)$$

es una medida de ajuste que puede ser aproximada utilizando la salida de la simulación MCMC de la distribución posterior, esta aproximación es dada por

$$Dbar = \frac{1}{G} \sum_{i=1}^G D(a^g, b^g, \lambda^g, \theta^g)$$

donde el índice  $g$  indica el  $g$ -ésimo valor simulado de un total de  $G$  simulaciones. El EAIC, EBIC y DIC pueden ser estimados de la siguiente manera

$$EAIC = Dbar + 2p$$

$$EBIC = Dbar + p \log N$$

$$DIC = Dbar + \rho_D = 2Dbar - Dhat \quad (9.10)$$

respectivamente donde  $p$  es el número de parámetros en el modelo,  $N$  es el total de observaciones,  $\rho_D$  es el número efectivo de parámetros y es definido como

---

<sup>15</sup> Estos dos últimos criterios generalmente no se los emplea, pueden ser calculados de manera analítica, pero aún están siendo evaluados por distintos autores de la filosofía Bayesiana.

$$\rho_D = E[D(a, b, \lambda, \theta)] - D[E[a], E[b], E(\lambda), E(\theta)]$$

Donde  $D[E[a], E[b], E(\lambda), E(\theta)]$  es el desvío de la media posterior obtenido cuando se evalúa la función desvío en la media posterior de los parámetros, el cual es estimado por

$$D\hat{h}at = D\left[\frac{1}{G} \sum_{i=1}^G a^g, \sum_{i=1}^G b^g, \sum_{i=1}^G \lambda^g, \sum_{i=1}^G \theta^g\right]$$

Para comparar dos o más modelos alternativos se usan los criterios *DIC*, *EAIC* y *EBIC*. En el *EAIC* y *EBIC*  $2p$  y  $plogN$  son valores fijos que penalizan a la media posterior del desvío. Desde que, no existe consenso en la literatura de que criterio sea el mejor, el uso de más de un criterio parece ser apropiado para realizar comparación de modelos.

*Dbar* y *DIC* son reportados en WinBUGS directamente cuando se requiere durante el proceso de simulación. *EAIC* y *EBIC* pueden ser derivados a partir del valor obtenido de *Dbar* considerando las expresiones presentadas.

El criterio de DIC asume que la media posterior es una buena estimativa para los parámetros del modelo. Así en casos en que las distribuciones son multimodales o en casos que existe una acentuada asimetría, el DIC no es recomendado.

DIC es semejante al AIC de enfoque clásico, mas difiere sustancialmente del criterio BIC o el Factor de Bayes.

## 10. INFORMACIÓN

La Encuesta Nacional de Demografía y Salud, principal fuente de información del País con relación al sector salud, es la más adecuada para realizar la presente investigación. En el módulo de “historia de nacimientos” de la ENDSA 2008 se obtiene información acerca de la fecha de nacimiento, edad actual, y edad al morir si es el caso, entre otros datos, para cada uno de los hijos nacidos vivos de las mujeres en edad fértil entrevistadas. A partir de esta información se elaboró el Cuadro 10.1, información que se usó en el análisis posterior.

### 10.1 ESTRUCTURA DE LA INFORMACIÓN

El Cuadro 10.1 expone las tasas de mortalidad neonatal (MNN), post-neonatal (MPNN) y de mortalidad post-infantil (MPI) para cada grupo de edad, cohorte de nacimiento, área de residencia y nivel de educación de las madres<sup>16</sup>. Se ha definido tres grupos de edad: 0 meses, 1-11 meses y 12-59 meses. Estos grupos de edad corresponden a las edades consideradas para calcular las tasas de mortalidad neonatal (MNN), post-neonatal (MPNN) y pos-infantil (MPI), respectivamente. Por esta razón, cuando por ejemplo se hace referencia a la tasa de MNN se estará haciendo mención a la mortalidad en el primer mes de vida. La variable cohorte también tiene tres categorías: la cohorte de nacidos vivos en el periodo 1993-1997, la cohorte de 1998-2002 y la cohorte de nacimientos en el periodo 2003-2007. Las variables residencia y educación tienen dos categorías: residencia urbana y rural para la variable lugar de residencia, y educación baja y alta para la variable nivel de educación. A madres sin educación formal o con educación primaria se las ha denominado

<sup>16</sup> Las estimaciones de mortalidad en las ENDSAS no son tasas sino probabilidades calculadas siguiendo los procedimientos estándar para construcción de tablas de mortalidad. Para cada período calendario se tabulan las muertes y las personas expuestas para los intervalos de edad en meses: 0, 1-2, 3-5, 6-11, 12-23, 24-35, 36-47 y 48-59, para luego calcular probabilidades de sobrevivencia en cada intervalo de edad. Finalmente se calculan las probabilidades de morir multiplicando las respectivas probabilidades de sobrevivir y restando de 1.

con educación baja, mientras a las madres con educación secundaria o superior se las ha denominado con educación alta.

Cuadro 10.1

Bolivia: Tasas de mortalidad observadas por tramo de edad, según cohorte de nacimiento, área de residencia y nivel de educación, ENDSA 2008

Área de residencia	Nivel de educación	Tramo de edad (en meses)	Cohorte		
			1993-1997	1998-2002	2003-2007
Urbana	Baja	0	0,3052	0,3622	0,3589
		1-11	0,0219	0,0240	0,0300
		12-59	0,0025	0,0061	0,0056
Urbana	Alta	0	0,1970	0,0240	0,2100
		1-11	0,0230	0,0196	0,0112
		12-59	0,0037	0,0015	0,0017
Rural	Baja	0	0,5731	0,6785	0,4655
		1-11	0,0433	0,0471	0,0240
		12-59	0,0074	0,0025	0,0064
Rural	Alta	0	0,2155	0,3125	0,3622
		1-11	0,0240	0,0219	0,0327
		12-59	0,0037	0,0240	0,0017

Fuente: Elaboración propia

\* Educación baja: Mujeres sin educación y con educación primaria

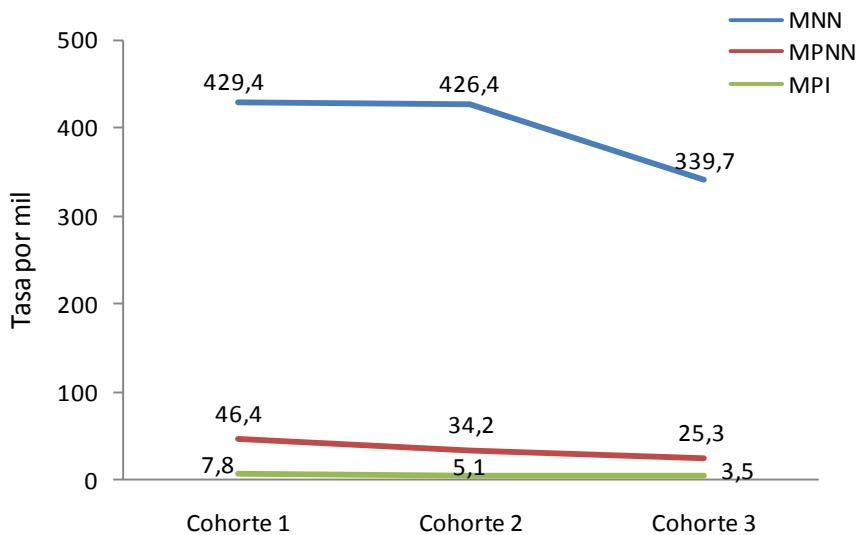
Educación alta: Mujeres con educación secundaria y con educación superior

A partir de la información en el Cuadro 10.1 puede analizarse las tasas de mortalidad neonatal (MNN), post-neonatal (MPNN) y de mortalidad post-infantil (MPI) para cada una de las tres cohortes de nacimiento en el capítulo 11 y capítulo 12<sup>17</sup>.

<sup>17</sup> Por la característica del estudio, se trabaja con probabilidades. Para más referencia ver “Indirect Techniques for Demographic Estimation”. Apéndice C.

Desde el punto de vista de un análisis descriptivo<sup>18</sup>, resaltan dos resultados en el Gráfico 10.1. Primero, el ritmo de descenso de la mortalidad difiere en cada grupo de edad. Esto es, la mortalidad neonatal parece haber descendido levemente entre la primera y la segunda cohorte, pero su descenso es notable al comparar la segunda con la tercera cohorte. En cambio, la mortalidad post-neonatal y la mortalidad post-infantil experimentaron ambos una constante reducción a lo largo del periodo de análisis. Segundo, se aprecia una amplia diferencia entre los niveles de los tres tipos de mortalidad, en cada una de las tres cohortes. En efecto, la mortalidad neonatal es entre 9 y 13 veces la mortalidad post-neonatal, y a su vez la mortalidad post-neonatal es entre 6 y 7 veces la mortalidad post-infantil. Estas últimas tasas – de mortalidad post-infantil – son muy bajas con relación a las otras tasas, lo cual refleja los logros en la reducción de la mortalidad post-infantil.

**Gráfico 10.1**  
**Bolivia: Tasas de mortalidad según cohorte de nacimiento, ENDSA 2008**



<sup>18</sup> Emplea solo las variables Edad y Cohorte de nacimiento en la información para realizar un análisis descriptivo.

## 11. RESULTADOS ENFOQUE CLÁSICO

Basado en los criterios del enfoque clásico se definió el modelo

$$EDA * COH + EDA * EDU + RESI$$

Cabe resaltar que este modelo elegido no tiene el problema de sobredispersión, pues el indicador de sobredispersión tiene un valor de uno.

### 11.1 Estimación del Modelo

Las estimaciones de los parámetros en el modelo de regresión Poisson elegido, y sus correspondientes errores estándar, están expuestas en el Cuadro 11.1.

Cuadro 11.1

Estimación de los parámetros (en términos de razón de tasas de incidencia) para el modelo "edad\*cohorte + edad\*educación"

Muertes	Razón de	Error Estándar	Intervalo de Confianza		
	tasas	(MIE)	z	P>z	(nivel 95%)
eda2	0,111	0,009	-27,22	0,000	0,095 0,130
eda3	0,019	0,002	-44,67	0,000	0,016 0,023
coh2	1,020	0,081	0,25	0,805	0,873 1,190
coh3	0,861	0,074	-1,73	0,083	0,726 1,020
edu2	0,527	0,045	-7,54	0,000	0,446 0,622
eda2coh2	0,745	0,084	-2,62	0,009	0,598 0,928
eda2coh3	0,697	0,088	-2,87	0,004	0,544 0,892
eda3coh2	0,675	0,085	-3,13	0,002	0,527 0,863
eda3coh3	0,584	0,102	-3,07	0,002	0,415 0,823
eda2edu2	0,843	0,107	-1,35	0,178	0,657 1,081
eda3edu2	0,638	0,104	-2,77	0,006	0,464 0,877

MIE: Matriz de información esperada

Para facilitar la interpretación, los efectos de las distintas variables están expresados en términos de razón de tasas de mortalidad<sup>19</sup>.

Para evaluar la sobredispersión del modelo se realizó una prueba de razón de verosimilitud para la hipótesis nula de que  $\alpha=0$ . La probabilidad de que exceda el valor 0 correspondiente a una chi-cuadrada con 1 grado de libertad es 0.5. De este modo se confirma que el modelo elegido, el cual fue estimado con una regresión Poisson, no tiene el problema de sobredispersión. En consecuencia, los errores estándar estimados y por tanto las inferencias basadas en éhos errores estándar, son correctos.

## 11.2 Predicción y Análisis de Residuos

En demasía importancia, es necesario evaluar el ajuste del modelo elegido comparando los valores predichos con los observados, para luego realizar el análisis de los coeficientes estimados. Esta comparación es útil para identificar algunas observaciones en las que el ajuste fue pobre. En el Gráfico 11.1 se comparan las muertes observadas con las predichas, mientras en el Gráfico 11.2 se hace la comparación en términos de las tasas de mortalidad.

En términos del número de muertes, el ajuste del modelo elegido a los datos en general parece razonable. Hay dos puntos, sin embargo, donde resaltan las diferencias entre el número de muertes observadas y predichas, ello ocurre cuando las muertes observadas son 114 y 187. Cuando la comparación es realizada en términos de tasas, el ajuste es bastante bueno en las primeras tres cuartas partes de las tasas, mientras en la última cuarta parte, vale decir en niveles relativamente elevados de mortalidad, la “calidad” del ajuste se reduce.

---

<sup>19</sup> Mantiene constante las demás variables.

Sólo con fines de comparación, en los Gráficos 11.3 y 11.4 se comparan el número de muertes y las tasas de mortalidad observadas, respectivamente, con las predichas usando otro modelo más complejo. Este modelo es más complejo que el modelo elegido por incluir en su parte sistemática, además de interacciones de segundo orden, interacciones de tercer orden, esto es, interacciones entre las variables edad, cohorte y educación. Por este hecho, este modelo aparentemente se ajusta mejor que el modelo elegido. Empero, los Gráficos 11.3 y 11.4 son muy parecidos a los Gráficos equivalentes 11.1 y 11.2. En términos del número de muertes, en los mismos dos puntos donde se registraron las mayores diferencias entre los valores observados y predichos con el modelo elegido también ocurren las mayores diferencias producidas con el modelo más complejo. Si la comparación es realizada en términos de tasas, la calidad de ajuste del modelo más complejo también es inferior en niveles relativamente elevados de mortalidad, igual a lo que se observó con el modelo elegido. En consecuencia, estos resultados indican que el modelo más complejo no es “mejor” que el modelo elegido. A esta misma conclusión se llegó con el test de razón de verosimilitud cuando se comparó ambos modelos.

Gráfico 11.1  
Número de muertes observadas y muertes predichas  
con el modelo elegido "EDA\*COH + EDA\*EDU + RES"

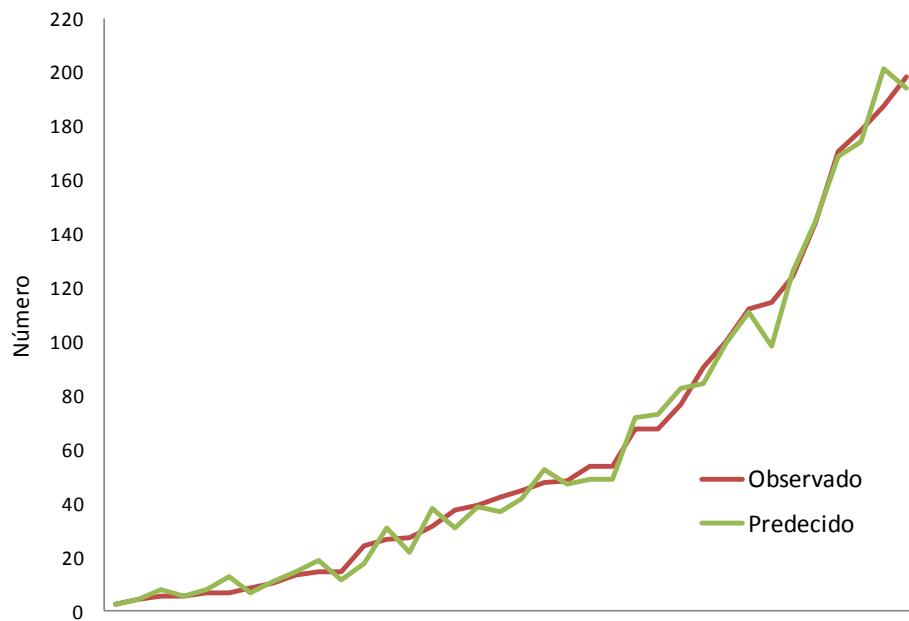


Gráfico 11.2  
Tasas de mortalidad observadas y tasas predichas  
con el modelo elegido "EDA\*COH + EDA\*EDU + RES"

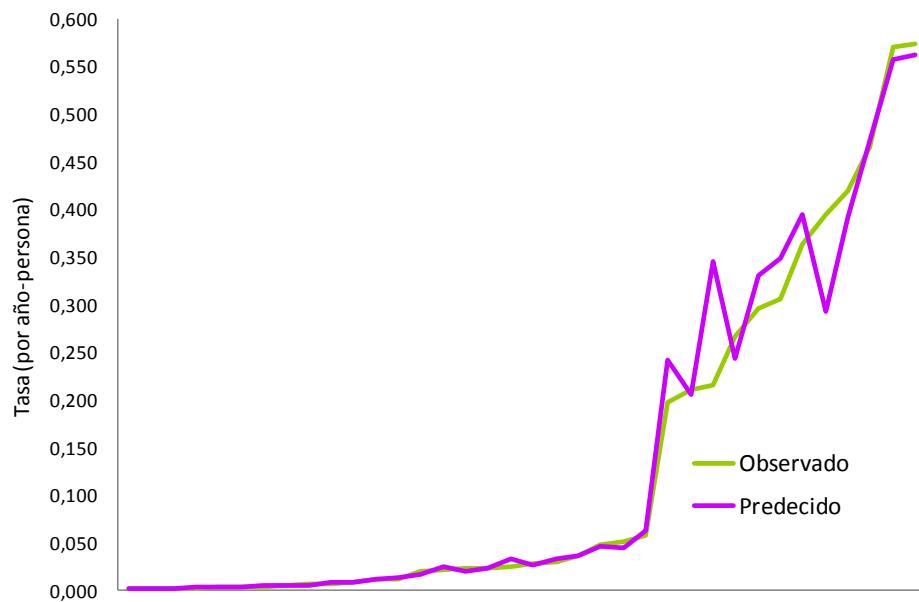


Gráfico 11.3  
Número de muertes observadas y muertes predichas  
con el modelo "EDA\*COH\*EDU + RES"

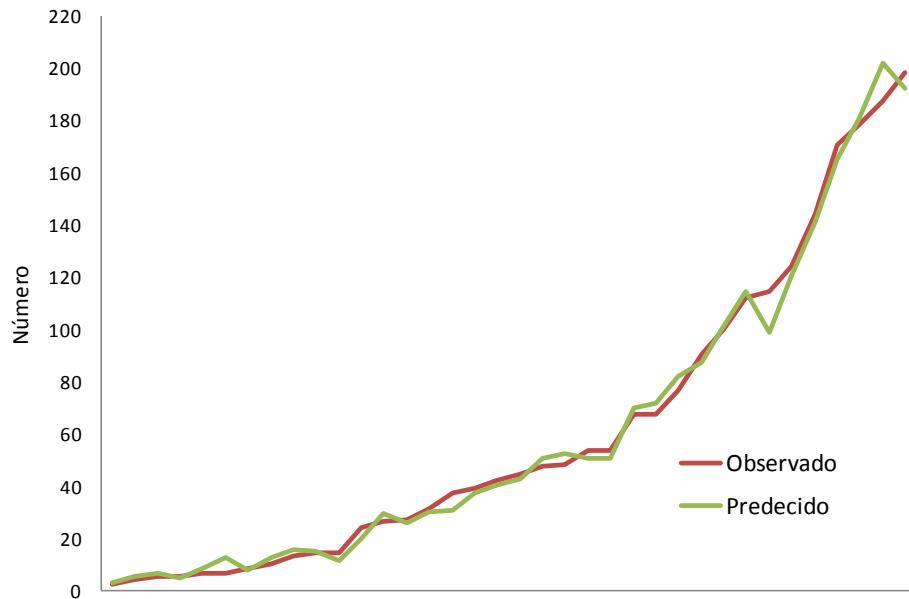
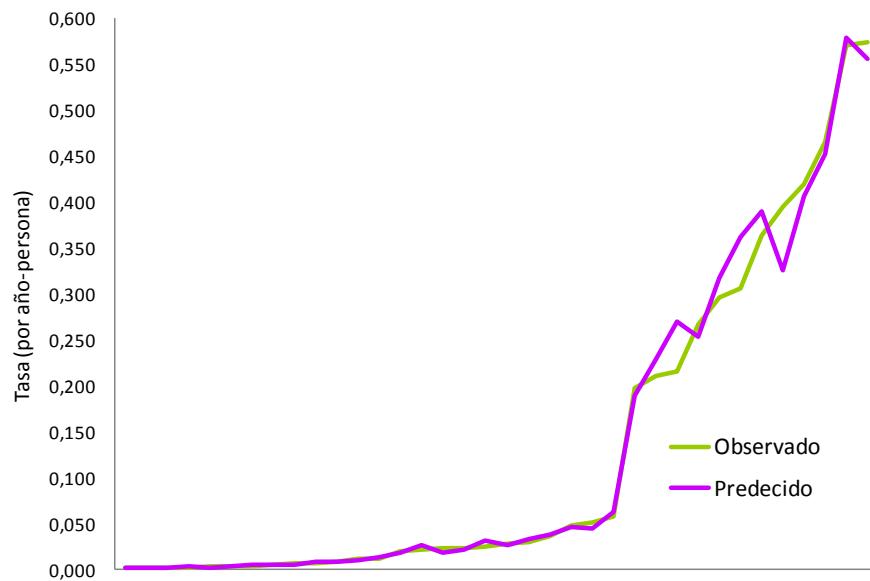


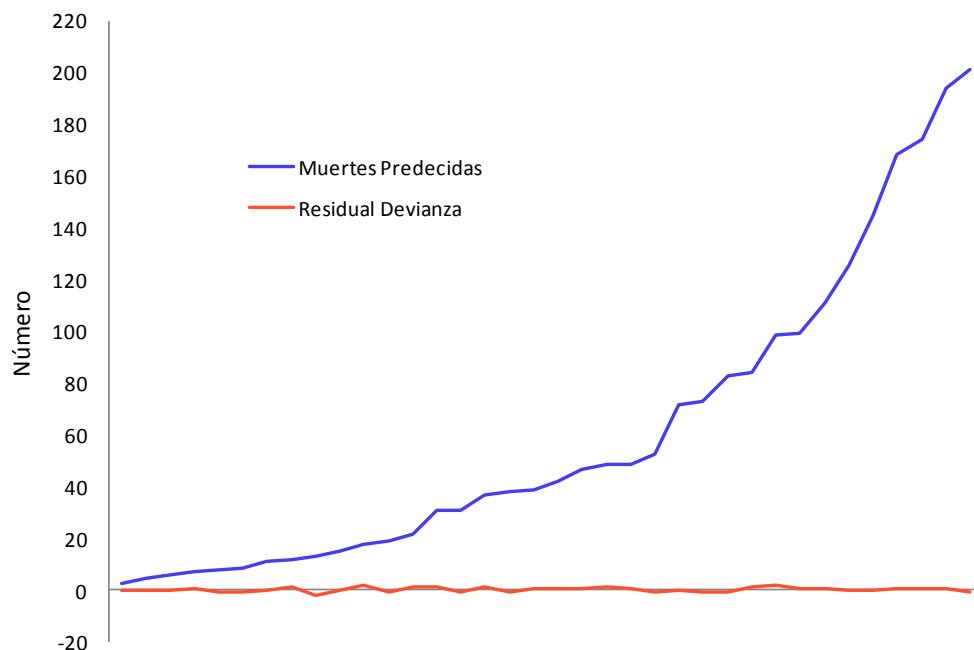
Gráfico 11.4  
Tasas de mortalidad observadas y tasas predichas  
con el modelo "EDA\*COH\*EDU + RES"



El análisis de los residuos proporciona una manera más fina de detectar las observaciones en las que el ajuste es pobre. Como se expuso en el acápite 8.3.2, el residual devianza está basado precisamente en el criterio de devianza. En el Gráfico 11.5 se exhiben estos residuos, junto con el número de muertes predichas con el modelo elegido.

Se puede apreciar en el gráfico que los residuos varían en torno de cero, sin desviaciones importantes de este valor. En teoría, si el modelo elegido es adecuado estos residuos tendrían que comportarse como una distribución normal. Para probar esta hipótesis se recurrió a los tests de Shapiro-Francia y Shapiro-Wilk. En ambos tests no se rechaza la hipótesis de normalidad. Con el test de Shapiro-Francia el valor-p es 0.884, mientras con el test de Shapiro-Wilk se obtuvo un valor-p de 0.941. Cabe hacer notar que la normalidad de los residuos derivados con el modelo complejo es rechazada al nivel de significancia de 5 por ciento cuando se usa el test de Shapiro-Francia. En consecuencia, estos resultados confirman que el modelo elegido describe adecuadamente las tasas de mortalidad observadas.

Gráfico 11.5  
Número de muertes predichas y residuos basados en el criterio de devianza para el modelo elegido "EDA\*COH + EDA\*EDU + RES"



### 11.3 Efectos y Significancias

En las secciones previas se evidenció que el modelo elegido se ajusta a los datos razonablemente bien y que no está afectado por el problema de sobredispersión. En lo que sigue se realiza un análisis más minucioso de los efectos y sus respectivas significancias.

#### **Efecto Cohorte**

Nuevamente, para determinar el efecto de la variable cohorte sobre la mortalidad se comparan las cohortes extremas (cohorte 1 y 3) por una parte, y las cohortes consecutivas (cohorte 1 y 2, y cohorte 2 y 3) por otra parte. En el

Cuadro 11.2 se muestran los efectos concernientes al primer caso, mientras los efectos correspondientes al segundo caso se presentan en el Cuadro 11.3.

Cuadro 11.2

Efecto cohorte: evolución de la mortalidad  
considerando las cohortes extremas, por  
tipo de mortalidad

Tipo de mortalidad	Coh3 vs. Coh1 (%)*)
MNN	-15,3 (0.0544)
MPNN	-41,0 (0.0000)
MPI	-50,5 (0.0000)

Fuente: Elaboración propia

\* En paréntesis el valor-p del efecto

El cuadro 11.2 presenta de manera explícita dos resultados. El primer resultado hace referencia a que las mortalidades post-infantil y post-neonatal experimentaron descensos que, en términos estadísticos, son altamente significativos en todo el periodo de análisis. Estas dos magnitudes de descenso son, estadísticamente, altamente significativas, pues sus respectivos valores-p son prácticamente 0. El segundo resultado indica que, a diferencia de los anteriores dos tipos de mortalidad, la mortalidad neonatal se mantuvo constante durante el periodo de análisis, pero no es estadísticamente significativa al nivel de 5 por ciento, pues su correspondiente valor-p es 0.0544.

Los resultados expuestos en el Cuadro 11.2 permitieron determinar la magnitud del descenso, y su significancia estadística, de cada tipo de mortalidad. En

cambio, los resultados del Cuadro 11.3 complementan el análisis anterior proporcionando elementos sobre el ritmo del descenso.

Cuadro 11.3

Efecto cohorte: evolución de la mortalidad considerando cohortes consecutivas, por tipo de mortalidad

Tipo de mortalidad	Cohortes Consecutivas	
	Coh2 vs Coh1 (%)*)	Coh3 vs Coh2 (%)*)
MNN	0,8 (0.9203)	-16,0 (0.0381)
MPNN	-24,9 (0.0016)	-21,5 (0.0042)
MPI	-31,9 (0.0004)	-27,3 (0.0127)

Fuente: Elaboración propia

\* En paréntesis el valor-p del efecto

El Cuadro 11.3 expone dos resultados (i) El ritmo de descenso tanto de la mortalidad post-neonatal como de la mortalidad post-infantil fue relativamente constante durante el periodo de análisis. (ii) En cambio, la evolución de la mortalidad neonatal es distinta a la de los otros dos tipos de mortalidad y, es más, para la mortalidad neonatal ahora surge un panorama en parte distinto al observado en el Cuadro 11.2. Por una parte, después de controlar el efecto de las variables educación y lugar de residencia, la mortalidad neonatal permanece constante entre la primera y la segunda cohorte, si bien se observa incluso un leve incremento de 0.8 por ciento, pero que estadísticamente no es significativa (valor-p de 0.9203). Por otra parte, ahora surge la evidencia de un descenso de la mortalidad neonatal entre la segunda y la tercera cohorte de 16 por ciento,

descenso que es significativo al nivel de significancia de 5 por ciento pero no al nivel de 1 por ciento, puesto que el correspondiente valor-p es 0.0381. Vale decir, la mortalidad neonatal habría descendido entre la segunda y tercera cohorte en una magnitud de 16 por ciento si se acepta un nivel de significancia de 5 por ciento; pero si se acepta un nivel de significancia de 1 por ciento la conclusión es que la mortalidad neonatal se habría mantenido constante entre la segunda y la tercera cohorte. El Gráfico 11.6 refleja la evolución de los tres tipos de mortalidad durante el periodo de análisis.

Sintetizando, después de controlar el efecto de las variables educación y lugar de residencia, tanto la mortalidad post-neonatal como la mortalidad post-infantil descendieron constantemente y significativamente durante el periodo de análisis, si bien el descenso para el segundo fue más importante que para el primero. Con éste resultado y con el obtenido cuando se analizó la información sólo con la edad y cohorte<sup>20</sup> (donde también se registró un descenso sistemático de ambos tipos de mortalidad) se concluye que la reducción tanto de la mortalidad post-neonatal como de la mortalidad post-infantil se debe al efecto conjunto de la educación, el lugar de residencia y de otros factores *no incluidos en el presente análisis*. Con relación a la mortalidad neonatal, una vez controlado el efecto de las variables educación y lugar de residencia, ésta se mantuvo constante durante el periodo de análisis. Nuevamente, con este resultado y con el obtenido cuando se analizó la información sólo con la edad y cohorte (donde se evidenció un descenso de la mortalidad neonatal, ocurrido principalmente entre las cohortes segunda y tercera) se concluye que en el periodo de análisis la mortalidad neonatal descendió significativamente y que tal reducción se debe *básicamente al efecto conjunto de la educación y de lugar de residencia*. Alguna evidencia estadística existe de que, además de la educación y del lugar de residencia, otros factores también pudieron haber contribuido al

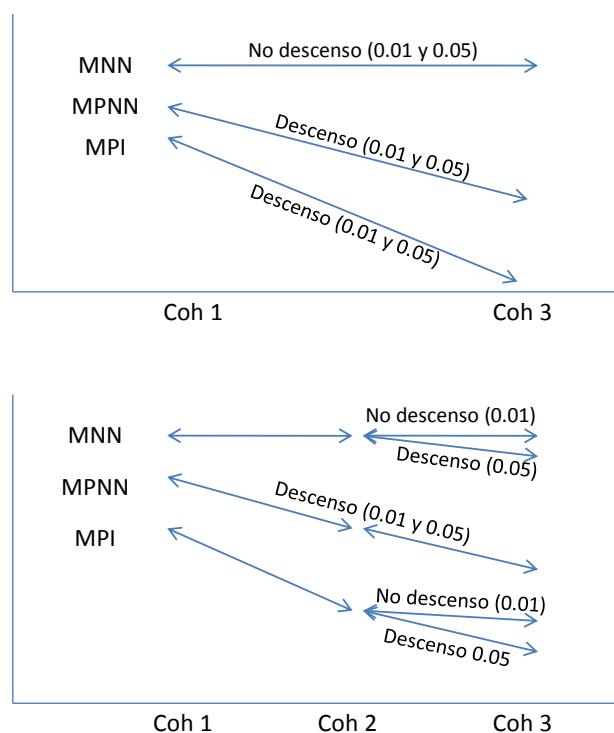
---

<sup>20</sup> Cuando sólo se toman las variables Edad y Cohorte de nacimientos (Información Agregada para fines descriptivos, que fue expuesta en los paneles para los tribunales, pero no se incluyó en el presente documento).

descenso de la mortalidad neonatal, pero en todo caso tal contribución es pequeña con relación a la contribución de la educación y del lugar de residencia (valor-p de 0.0381).

Gráfico 11.6

Tendencia de la mortalidad en los primeros años de vida,  
una vez controlado el efecto de educación y residencia



## 12. RESULTADOS ENFOQUE BAYESIANO

Para el mismo modelo definido bajo el enfoque clásico se procedió a estimar los parámetros con el enfoque bayesiano

### 12.1 Estimación del Modelo

Las estimativas de los parámetros del modelo elegido mediante el enfoque Bayesiano se muestran en el Cuadro 12.1. Tales estimaciones se desarrollaron en base a distribuciones previas no informativas para los parámetros. Se desechó las primeras 1000 burn – in para que la cadena olvide su estado inicial.

Cuadro 12.1

Estimación bayesiana de los parámetros para el modelo elegido "EDA\*COH + EDA\*EDU + RES"

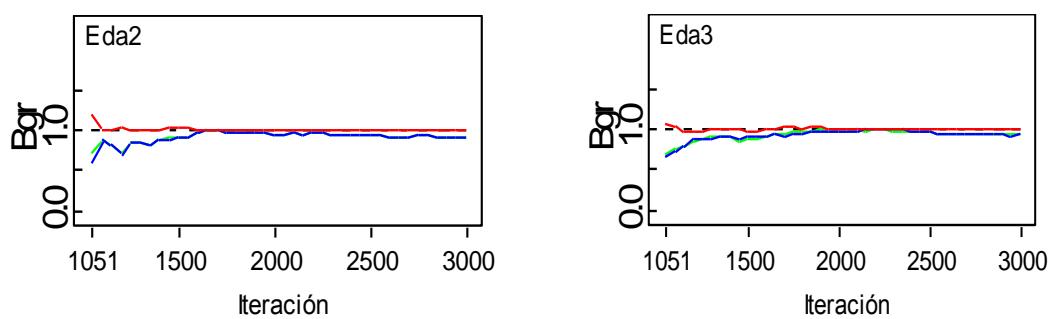
node	mean	sd	MC error	2,50%	median	97,50%	start	sample
eda2	-2,195	0,077	0,005	-2,348	-2,196	-2,040	1001	4000
eda3	-3,946	0,084	0,004	-4,108	-3,946	-3,785	1001	4000
coh2	0,011	0,076	0,004	-0,141	0,011	0,162	1001	4000
coh3	-0,168	0,083	0,004	-0,326	-0,168	0,000	1001	4000
edu2	-0,481	0,088	0,003	-0,654	-0,482	-0,311	1001	4000
res2	-0,356	0,048	0,001	-0,449	-0,357	-0,261	1001	4000
eda2coh2	-0,298	0,108	0,006	-0,504	-0,297	-0,084	1001	4000
eda2coh3	-0,361	0,121	0,006	-0,608	-0,359	-0,129	1001	4000
eda3coh2	-0,399	0,124	0,005	-0,637	-0,401	-0,151	1001	4000
eda3coh3	-0,546	0,169	0,005	-0,879	-0,545	-0,217	1001	4000
eda2edu2	-0,171	0,125	0,004	-0,420	-0,172	0,071	1001	4000
eda3edu2	-0,456	0,162	0,005	-0,786	-0,452	-0,151	1001	4000
cons	-0,587	0,057	0,004	-0,698	-0,587	-0,473	1001	4000

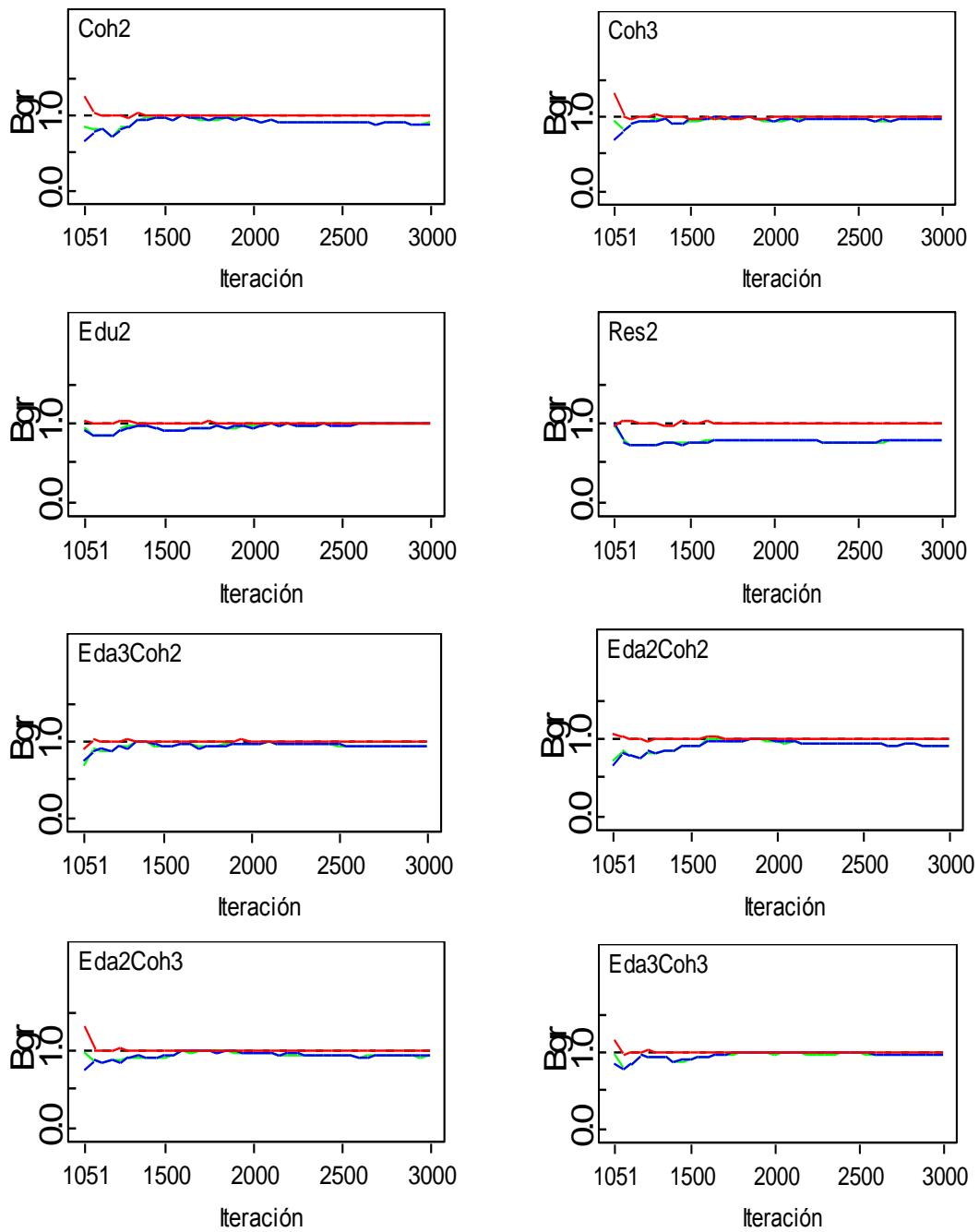
El cuadro precedente presenta la media y el desvío de la distribución, así como el intervalo de probabilidad al 95% para cada uno de los parámetros. Tales estimaciones son muy similares a las obtenidas por el enfoque Clásico.

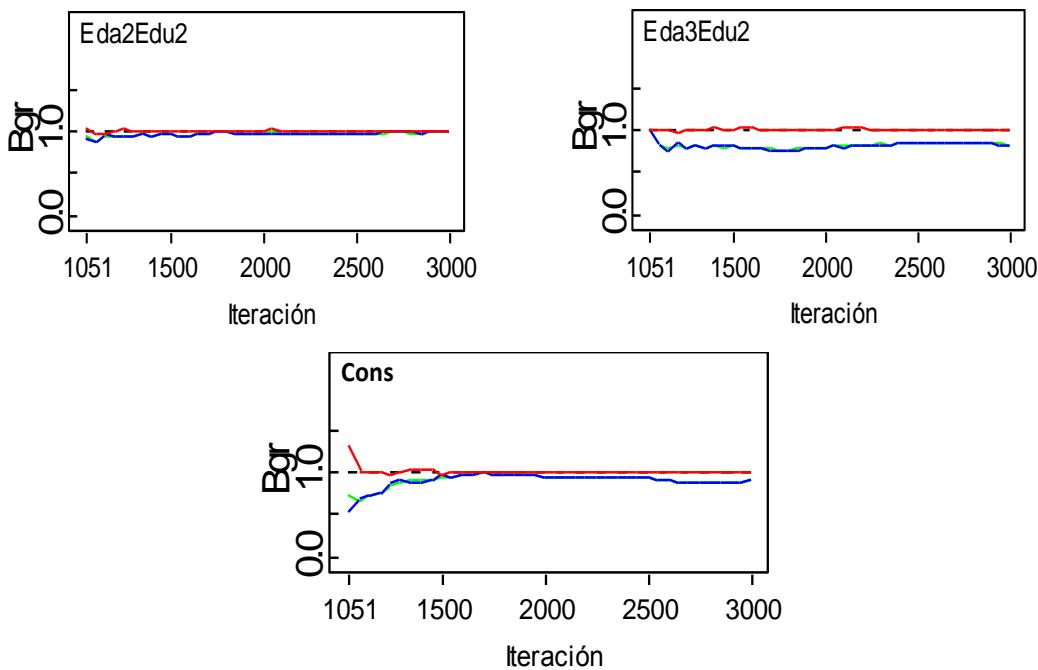
## 12.2 Diagnóstico de Convergencia

Se corrieron 5000 iteraciones, eliminando las 1000 primeras. El estadístico de Gelman y Rubin observado indica que se obtuvo la convergencia deseada, esto permite confiar en que los resultados obtenidos refieren a las distribuciones posteriores buscadas. En el Gráfico 12.1 se presentan el estadístico de Gelman y Rubin para cada uno de los parámetros del modelo elegido.

Gráfico 12.1  
Diagnóstico de Convergencia para cada uno de los parámetros  
Modelo Eda<sup>\*</sup>Coh+Eda<sup>\*</sup>Edu+Res



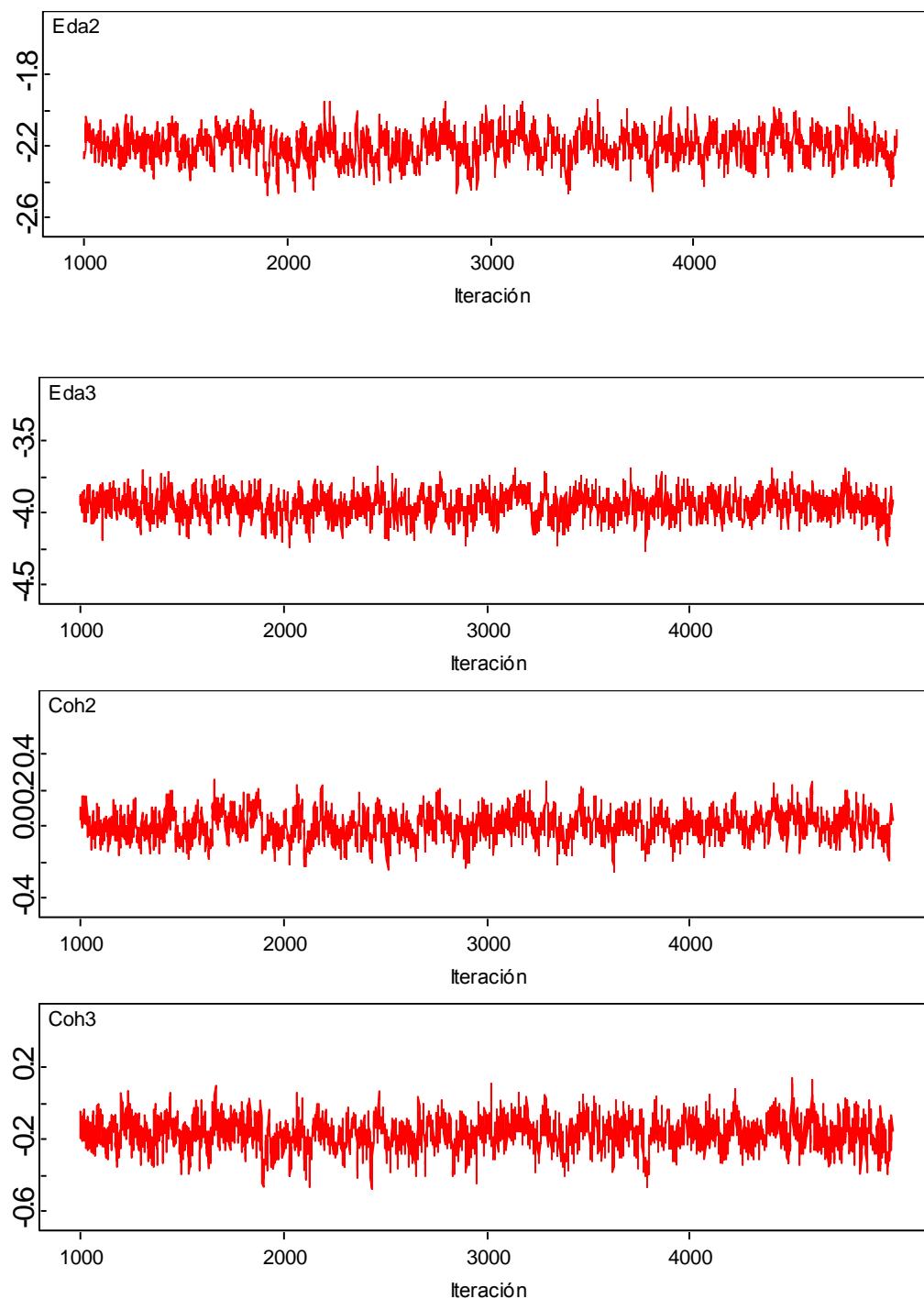


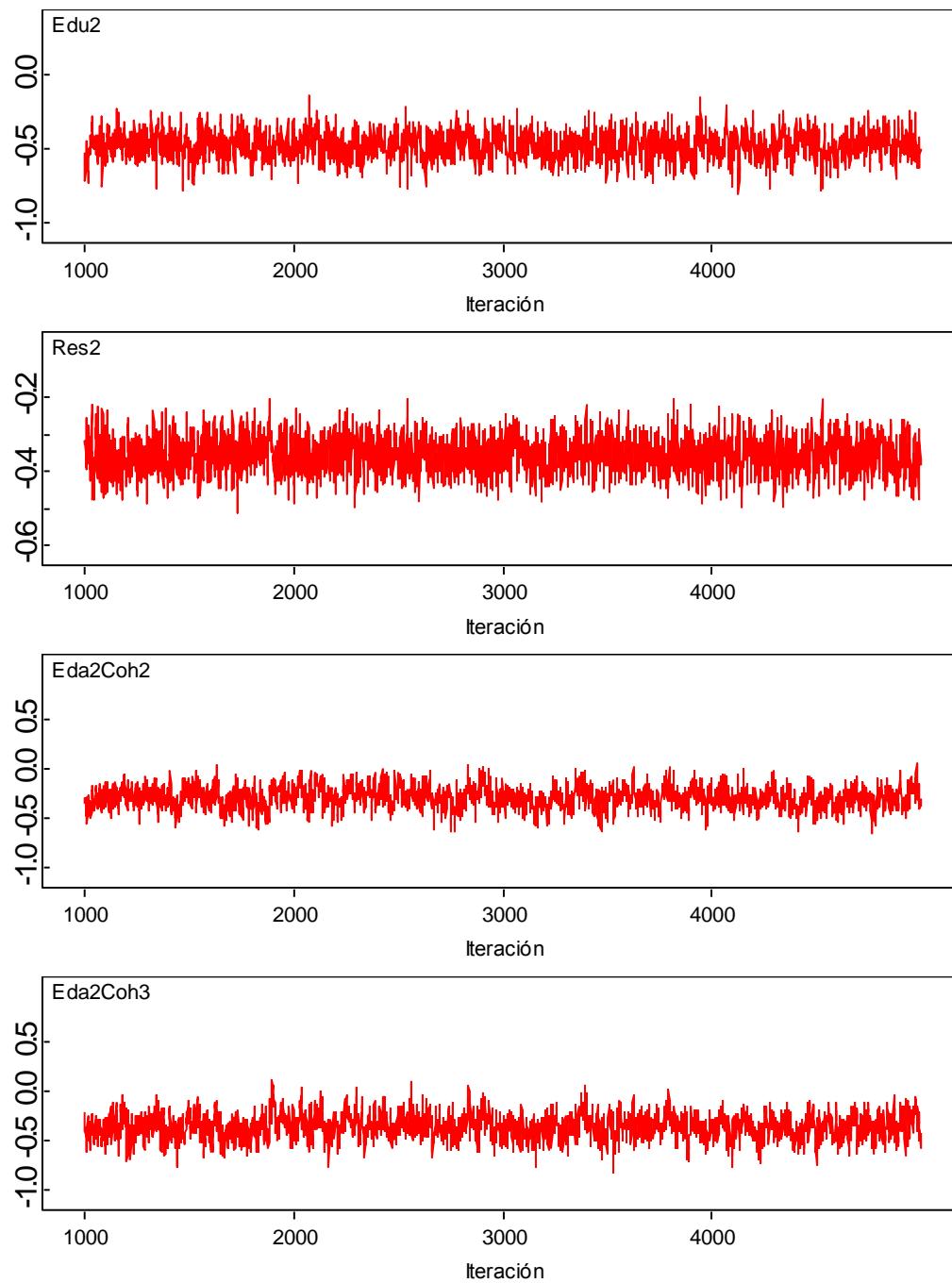


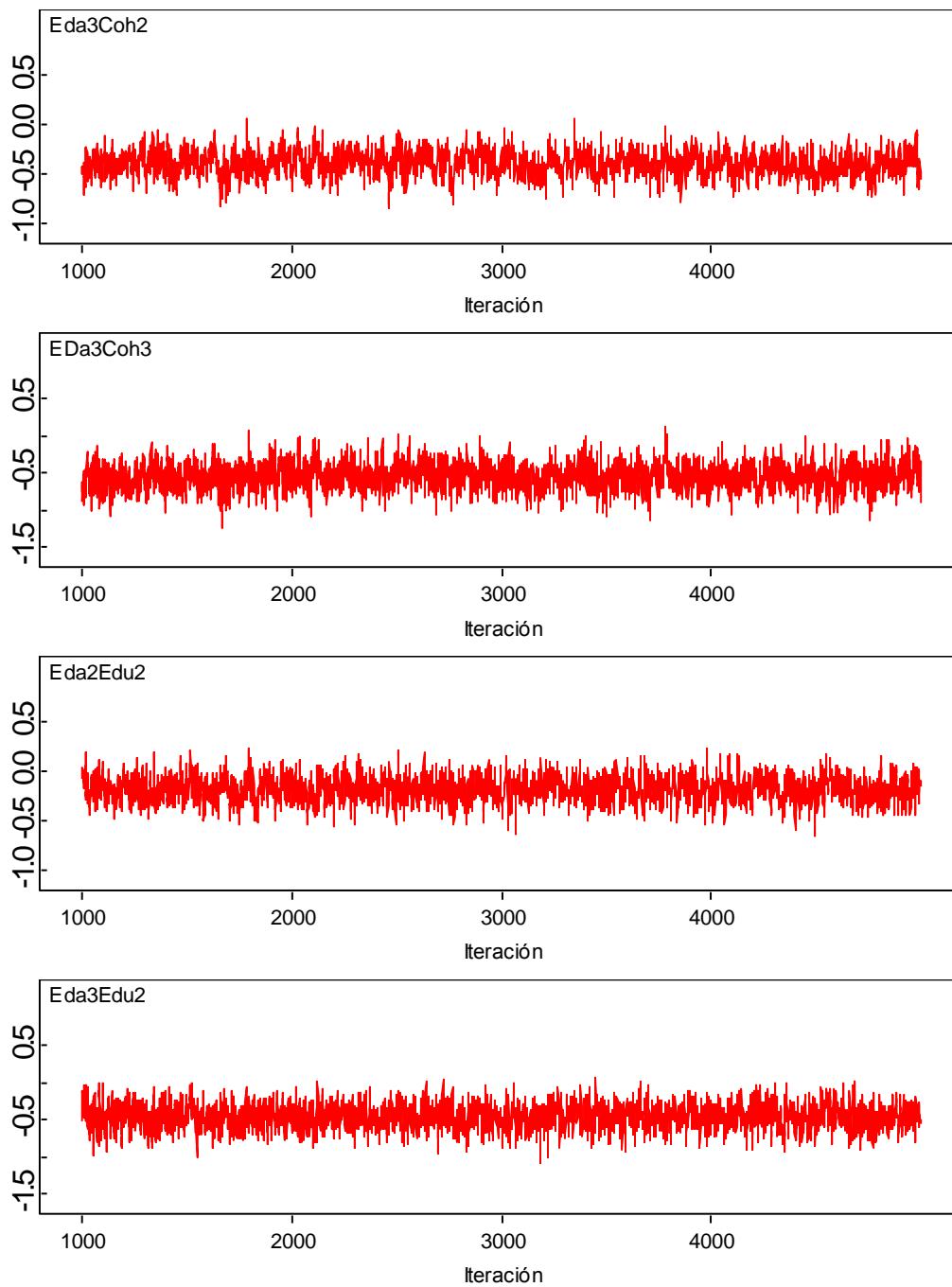
Al observar los gráficos se puede afirmar que las estimaciones de la distribución de los parámetros alcanzaron la convergencia. Se aprecia que el estadístico (representado con color rojo en el gráfico) se encuentra en el entorno de 1, como era deseable. Los colores verde y azul son representaciones de las varianzas estimadas y la varianzas entre cadena, respectivamente.

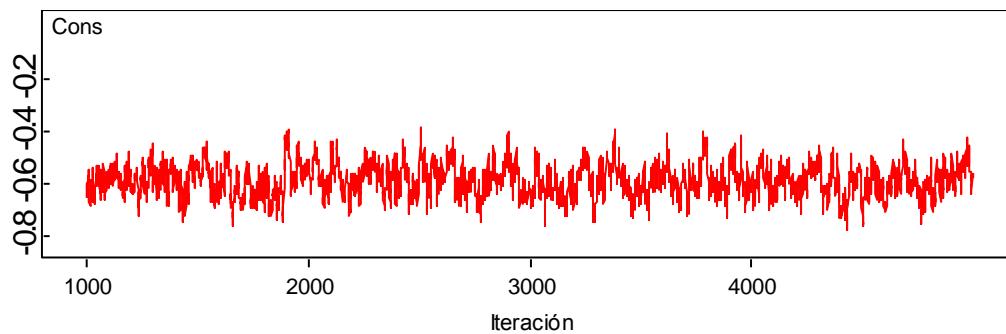
Para mayor convicción, en el Gráfico 12.2 se observa las historias de cada parámetro y su estabilización en el transcurso de las 5000 iteraciones. Se observa la convergencia hacia su distribución posterior.

Gráfico 12.2  
Historias de la convergencia de los parámetros  
Evolución de la Simulación



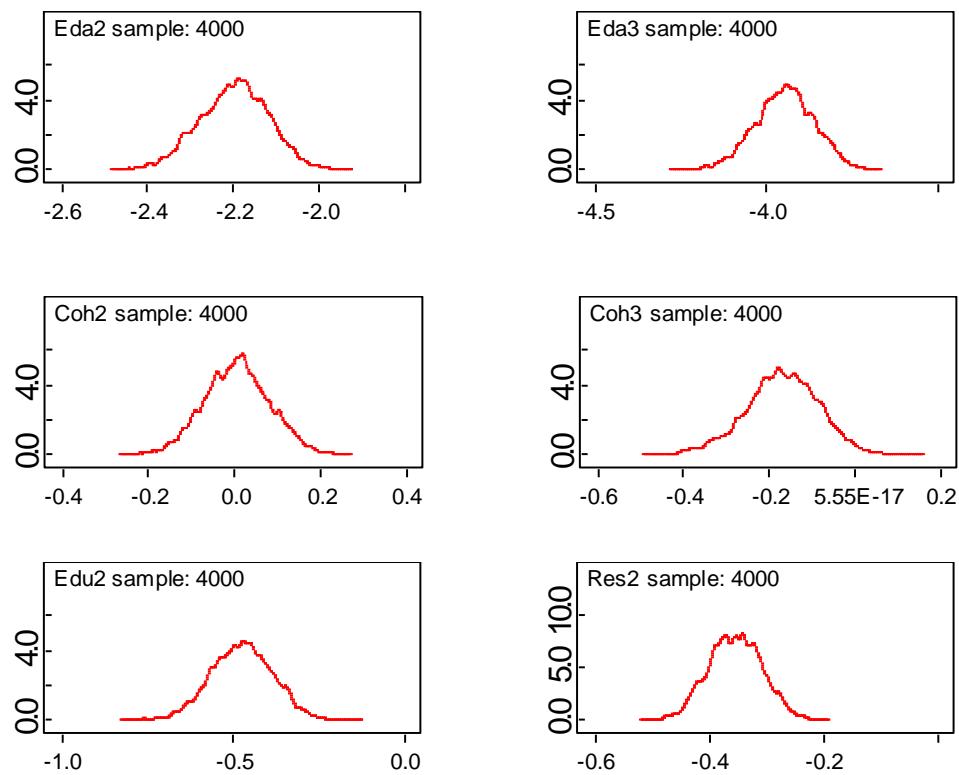


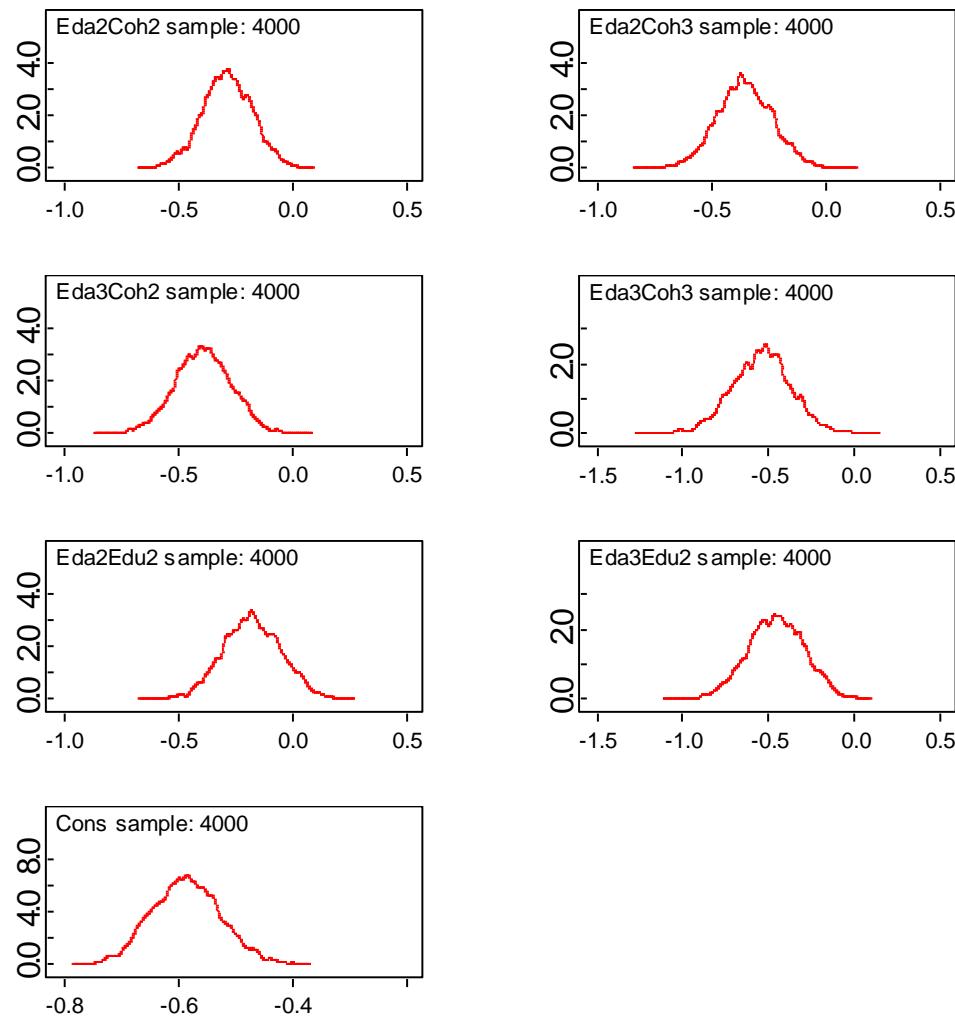




Otra opción visual adicional es la expresa el Gráfico 12.3, se evidencia las densidades de Kernel de la distribución posterior de los parámetros.

Gráfico 12.3  
Densidad estimada de la Distribución Posterior





Se observa una distribución unimodal simétrica para cada uno de ellos, (lo cual también se puede verificar analíticamente en el Cuadro 12.1 de las distribuciones estimadas ya que la media y la mediana prácticamente coinciden).

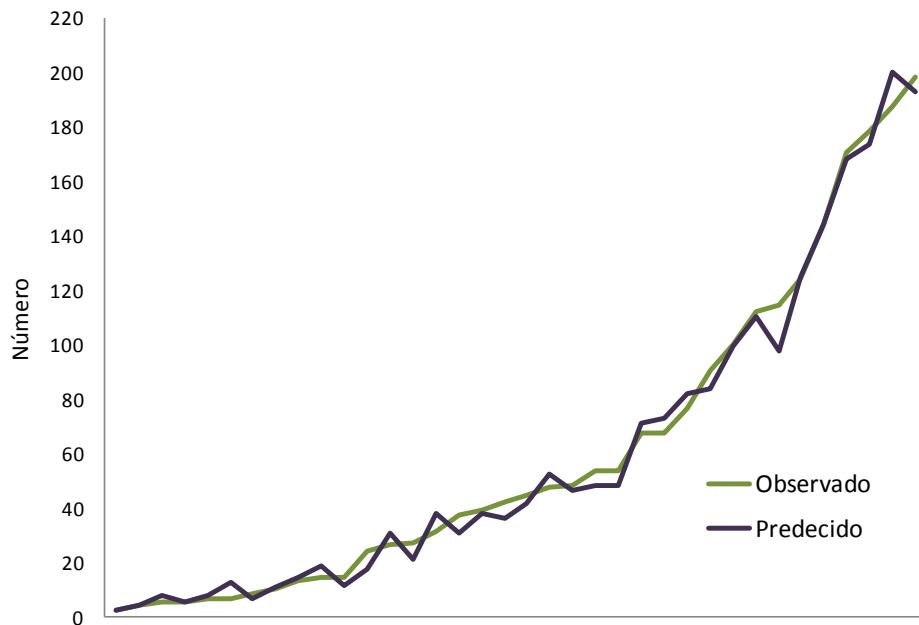
### 12.3 Predicción

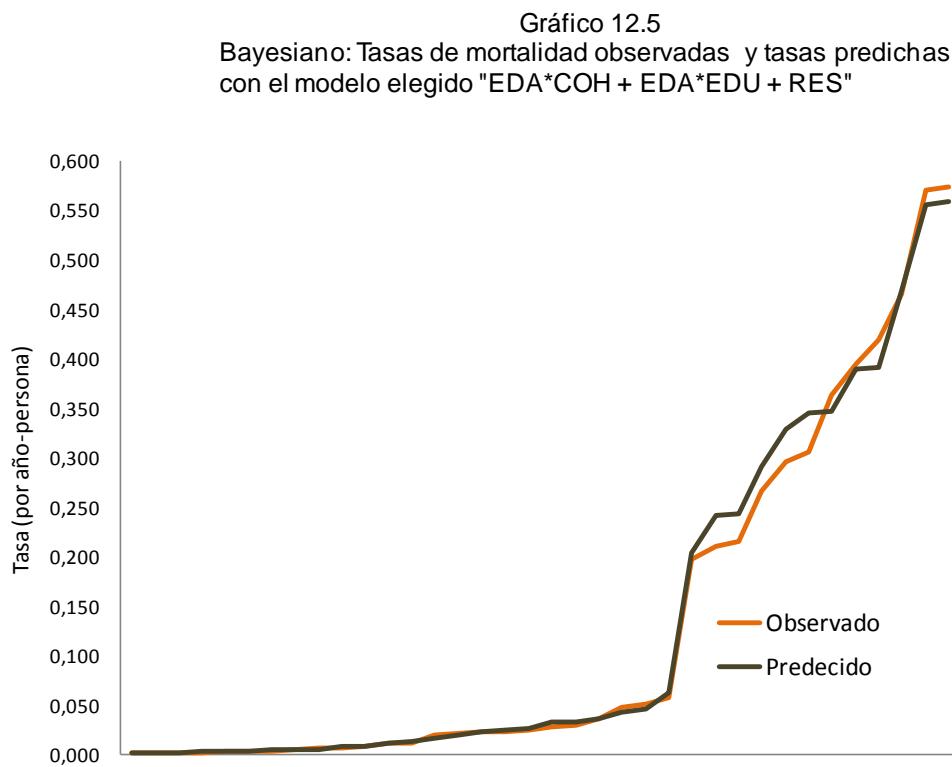
Es necesario evaluar el ajuste del modelo elegido comparando los valores predichos con los valores observados para identificar algunas observaciones en

las que el ajuste fue pobre. En el Gráfico 12.4 se comparan las muertes observadas con las predichas, mientras en el Gráfico 12.5 se hace la comparación en términos de las tasas de mortalidad, ambos para el modelo elegido.

Como se observó en el análisis clásico, el Gráfico 12.4 es igual al Gráfico 11.1 en términos del número de muertes. Hay dos puntos, sin embargo, donde resaltan las diferencias entre el número de muertes observadas y predichas, ello ocurre cuando las muertes observadas son 114 y 187. Cuando la comparación es realizada en términos de tasas, el ajuste es bastante bueno y mejor que en el análisis clásico, también en la última cuarta parte la “calidad” del ajuste se reduce.

Gráfico 12.4  
Bayesiano: Número de muertes observadas y muertes predichas  
con el modelo elegido "EDA\*COH + EDA\*EDU + RES"





## 12.4 Efectos

En lo que antecede se vio que el modelo se ajusta razonablemente a los datos y que el monitoreo de convergencia resultó ser aprobado, resta hacer un análisis minucioso de los efectos.

### Cohorte

En el Cuadro 12.2 se puede evidenciar resultados similares a los resultados del enfoque clásico, comparando cohortes extremas se observa que en el periodo de análisis (2003 – 2007), la mortalidad neonatal experimentó un descenso de 15.7 por ciento, la mortalidad post-neonatal descendió 40.8 por ciento y la mortalidad post-infantil 50.9 por ciento.

Cuadro 12.2

Bayesiano  
Efecto Cohorte: descenso de la mortalidad  
considerando Cohortes extremas  
por tipo de Mortalidad

Tipo de Mortalidad	Cohorte	
	1993-1997	2003-2007 (%)
0	-	-15,7
1-11	-	-40,8
12-59	-	-50,9

En el Cuadro 12.3 se expone el descenso de la mortalidad considerando cohortes consecutivas. Se puede evidenciar similares resultados a los del enfoque Clásico,

Cuadro 12.3

Bayesiano  
Efecto Cohorte: descenso de la mortalidad  
considerando Cohortes consecutivas  
por tipo de Mortalidad

Tipo de Mortalidad	Cohorte		
	1993-1997	1998-2002 (%)	2003-2007 (%)
0	-	0,5	-16,1
1-11	-	-24,6	-21,5
12-59	-	-32,1	-27,8

### 13. CONCLUSIONES Y RECOMENDACIONES DE POLÍTICA

En el periodo de análisis, 1993 a 2007, los tres tipos de mortalidad - neonatal, post-neonatal y post-infantil - experimentaron descensos, las dos últimas estadísticamente significativos, pero la MNN experimentó un descenso estadísticamente no significativo. Tales descensos difieren tanto en magnitud como en su ritmo. La tasa de mortalidad neonatal se redujo en sólo 15.3 por ciento, la tasa de mortalidad post-neonatal en 41.0 por ciento, mientras la tasa de mortalidad post-infantil lo hizo en 50.5 por ciento. Como se indicó, igualmente difieren sus ritmos de descenso. La mortalidad post-neonatal y la post-infantil se redujeron aproximadamente en forma continua durante el periodo de análisis, en cambio la mortalidad neonatal permaneció constante entre la primera y la segunda cohorte de nacimientos y su descenso fue estadísticamente significativo se registra entre la segunda y tercera cohorte.

Producto de las diferencias en cuanto a magnitud y ritmo de descenso, el orden de importancia de los tres tipos de mortalidad se alteró. Esto es, entre los niños de la primera cohorte, vale decir entre los nacidos en 1993-1997, la mortalidad post-neonatal concentra el 38.8 por ciento de las muertes ocurridas en los primeros cinco años de vida, seguido por la mortalidad neonatal (34.1 por ciento) y por último la mortalidad post-infantil (27.1 por ciento). Diez a quince años después esta distribución se altera. Entre los niños de tercera cohorte, es decir entre los nacidos en 2003-2007, la mortalidad neonatal pasa a concentrar el 50.6 por ciento de las muertes, la mortalidad post-neonatal ahora aglutina el 38.3 por ciento y el restante 11.1 por ciento corresponde a muertes post-infantiles. No cabe duda que las políticas relacionadas con la reducción de la mortalidad en la niñez deberían enfocarse principalmente en evitar las muertes neonatales, puesto que la mayoría de las muertes en los primeros cinco años

de vida ahora ocurren en el primer mes de vida, sin desmerecer las acciones orientadas a erradicar la mortalidad post-neonatal y post-infantil.

A fin de determinar el efecto de las variables nivel de educación de las madres y área de residencia en el descenso de los tres tipos de mortalidad, estas dos variables se incluyeron en la parte sistemática del modelo de regresión Poisson. Con base en el modelo elegido (el cual contiene un término de interacción entre la edad y la corte de nacimientos y otro entre la edad y el nivel de educación, más la variable área de residencia) se encuentra evidencia estadística de que el descenso de la mortalidad neonatal en el periodo de análisis se debe *básicamente al efecto conjunto de la educación y de lugar de residencia*. Alguna evidencia estadística existe de que, además de la educación y del lugar de residencia, otros factores también pudieron haber contribuido al descenso de la mortalidad neonatal, pero en todo caso tal contribución es pequeña con relación a la contribución de la educación y de lugar de residencia. En cambio, la reducción tanto de la mortalidad post-neonatal como de la mortalidad post-infantil se debe al efecto conjunto de la educación, lugar de residencia y *de otros factores no incluidos en el presente análisis*.

Estas conclusiones son coherentes con lo que la literatura sobre el tema expresa. Por una parte, una reducción en la mortalidad neonatal requiere particularmente de adecuado diagnóstico prenatal y de atención especializada en el período perinatal puesto que la primera semana de vida es la que más riesgo entraña para los recién nacidos. En el país, la atención postnatal tanto a madres como a sus recién nacidos aun está limitada a ciertos grupos de población, principalmente a los que residen en áreas urbanas y a los de mayor nivel educativo o mayor ingreso. En efecto, según resultados de la END SA 2008, el 83.4 por ciento de las madres con educación superior recibió el primer control postnatal para el nacimiento más reciente en el primer día después del parto, frente a sólo un 38.8 por ciento en el caso de madres sin educación.

Aunque un poco menos amplias pero igualmente importantes son las diferencias entre los porcentajes de atención postnatal de áreas y urbanas y rurales. Por otra parte, la mortalidad post-neonatal y post-infantil puede evitarse mediante intervenciones médicas y sanitarias. Por ejemplo, la lactancia materna, la terapia de rehidratación oral y la vacunación pueden reducir significativamente ambos tipos de mortalidad. En el país, acciones de esta naturaleza fueron el principal contenido de los programas relacionados a la prevención de la mortalidad en los primeros años de vida.

Si bien ambas variables – nivel de educación y área de residencia – contribuyeron significativamente tanto al ritmo como a la magnitud de descenso de los tres tipos de mortalidad, la contribución de la educación fue mayor que la de residencia. Esto es, la tasa de mortalidad neonatal en hijos de madres con educación alta es 38.1 por ciento menos que en hijos cuyas madres tienen educación baja; la tasa de mortalidad post-neonatal en el grupo de educación alta es 47.8 por ciento menos que en el de educación baja; mientras los hijos de madres más educadas tienen una tasa de mortalidad post-infantil 60.4 por ciento menos que los hijos de madres menos educadas. Como se puede ver, el efecto de la educación en la mortalidad depende de la edad en la que ocurren las muertes. En cambio, el efecto del lugar de residencia en la mortalidad, aunque igualmente importante, es inferior al de la educación. Esto es, la tasa de mortalidad en el área urbana es 29.9 por ciento menos que en el área rural. Este efecto es el mismo en cada categoría de educación (baja y alta), en cada cohorte de nacimientos (primera, segunda y tercera) y para cada tipo de mortalidad (neonatal, post-neonatal y post-infantil).

En el orden metodológico, las estimaciones de los parámetros que se obtuvieron con ambos enfoques (Clásico y Bayesiano) resultaron ser las mismas, produciendo conclusiones iguales en el momento de responder a los

objetivos. El problema de sobredispersión que se tuvo inicialmente con el modelo de regresión Poisson fue aparente, no real<sup>21</sup>.

Al analizar la información con la metodología Bayesiana se tornó complicado, pues al seguir la metodología que profana, existió una exigencia de un amplio conocimiento de teoría estadística Clásica, programación, y por último (en este trabajo de tesis), nociones básicas en Demografía. Aún así, los avances computacionales también juegan un papel en contra cuando no es posible encontrar softwares (o si los hay, trabajan como “caja negra”) que se estructuren sobre métodos bayesianos ya definidos, pues hoy en día la metodología Bayesiana continúa evaluando sus métodos (desechando y creando nuevos). En síntesis, los métodos bayesianos son más complejos que los clásicos, y el software disponible es escaso, pero cuando la metodología Bayesiana funciona, el resultado del trabajo es magnífico y gratificante.

---

<sup>21</sup> Aclárese que todos los resultados presentados en ésta tesis, fueron fundamentados explícitamente mediante programación en vivo en los paneles previos presentados antes de la presentación final.

## 14. APÉNDICES

### Apéndice A

#### *Programas empleados para analizar la información*

Para la manipulación de la información tanto en el análisis clásico como en el análisis Bayesiano se eligió en primera instancia al programa “R 2.11.0” por su versatilidad al trabajar en distintas áreas de la estadística,

Para contrastar los resultados del análisis Bayesiano obtenidos con el programa R, también se analizó la información con los programas WinBugs y OpenBugs llegando a los mismos resultados. Además alternativamente se usó las interfaces R2WinBUGS y R2BRugs de “R 2.11.0” con Bugs

Dentro de R, se cargó la librerías R2WinBUGS y R2BRugs que fueron instalados previamente library(R2WinBUGS), library(R2BRugs), además se verifico la existencia de las library's “coda” y “lattice”, de lo contrario no corre.

### Apéndice B

A continuación se presenta los programas necesarios para implementar la metodología propuesta en este trabajo de investigación. Además, estos programas permiten llevar a cabo y reproducir los resultados obtenidos. Los programas Programa con BRugs.R, metrópolis hastingR, muestreador Gibbs y iteraciones corren en R con sus librarys correspondientes.

```
# EXPORTAR DATOS DE STATA outfile using "E:/PATRICIA/R/REGRESION  
POISSONBAYESIANO/INFORMACION/TASASDESAGREGADO.raw"  
# RECUPERAR DATOS EN R matrix(scan("E:/PATRICIA/R/REGRESION  
POISSONBAYESIANO/INFORMACION/TASASDESAGREGADO.raw"), nrow=36, ncol=13, byrow=TRUE)  
  
salida<-bugs(data,inits,parameters.to.save=matrix ("beta","lambda"),
```

```
model.file="E:/PATRICIA/modeloelegido.txt", n.chains=2, n.iter=8000,  
n.burnin=1000,program="BRugs")
```

Alternativamente, R tiene una librería basada en la versión OpenBugs que funciona como cualquier otra librería de R: BRugs.

El programa correspondiente para R es:

```
rm(list=ls(all=TRUE)) # Limpia la memoria  
  
ruta <- "E:/PATRICIA/"  
setwd(INFORMACION) # Fija la ruta de trabajo  
library(BRugs)  
  
# Programa con BRugs  
  
modelo <- BRugsFit(data=datos, inits=iniciales,  
para=parametros, nBurnin=5000, nIter=10000,  
modelFile="modelo.txt",  
numChains=3, working.directory=ruta)  
pSamp <- samplesSample("p")  
pSamp # Resultados de la Estimación  
pMed <- median(pSamp)  
cat("La mediana de las simulaciones es:", pMed, "\n")  
X11()  
plotDensity("p", xlab="p", main="Densidad de p")  
X11()  
histinfo <- hist(pSamp, xlab="p", ylab="Frecuencias",  
freq=F, main="p estimada", col="lightblue")  
file.remove("modelo.txt")  
  
# Programa con BRugs  
modelo <- BRugsFit(data=datos, inits=iniciales,  
para=parametros, nBurnin=5000, nIter=10000,  
modelFile="modelo.txt",  
numChains=3, working.directory=ruta)  
samplesStats("*")  
png(file="cosa.png", pointsize=8)  
par(mfrow=c(2,1))  
plotDensity("theta1", xlab=expression(theta), main="Densidad de theta1")  
plotDensity("theta2", xlab=expression(theta), main="Densidad de theta2")  
dev.off()  
file.remove("modelo.txt")  
  
#### ----- Iteraciones -----  
for(k in 1:kk)  
{  
  flag1<-(k/1000)-trunc(k/1000)  
  if(flag1==0){print(k)}  
  # Sampling of vectors beta1 and beta2.
```

```

XtY1<-interacprod (X.1,Y[,1])
XtY2<- interacprod (X.2,Y[,2])
mstar1<-c interacprod (Sigma1,((L1m1)+XtY1) )
mstar2<-c(interacprod (Sigma2,((L2m2)+XtY2) ))
beta1<-mvrnorm(1,mstar1,Sigma1)
beta2<-mvrnorm(1,mstar2,Sigma2)
# Sampling of vector r
for(j in 1:n)
{
t.aux<-data$theta[j]
mu.b1<-c(interacprod (beta1,X.1[j,]))
mu.b2<-c(interacprod (beta2,X.2[j,]))
r[j]<-Metro(lirt,t.aux,mu.b1,mu.b2,1,5)
}
Y<-r*datose
#----- Valores de cada iteracion -----
flag<-(k/t.lag)-trunc(k/t.lag)
if(flag==0)
{
ii<-k/t.lag
B1[ii,<-beta1 B2[ii,<-beta2 ]
#----- Termina el algoritmo de Gibbs -----
}
#
#-----#
# Calculo de la medida predictiva Lk. Lk<-0.0
flag.LK<-ifelse(flag.lk=="TRUE",1,0) if(flag.LK){ mu<-matrix(0,tm,2)
predictiva<-rep(0,n) for(i in 1:n){
mu[,1]<-c(crossprod(t(B1),X.1[i,]))
mu[,2]<-c(crossprod(t(B2),X.2[i,])) norm2.mu.i<-norm2.row(mu)
vtnmu.i<-c(crossprod(t(mu),datose[i,]))
predic.k<-(1/(2*pi))*exp(-0.5*norm2.mu.i)
*( 1 + ((vtnmu.i*pnorm(vtnmu.i))/dnorm(vtnmu.i)) )
predictiva[i]<-mean(predic.k)
}
Lk<-prod(predictiva)
}
#
#-----#
# Salida:
B<-list(BI=B1,BII=B2,Lk=Lk)
drop(B)
}
#
#-----#
#
#
#
```

```

Dbd<-function(t,mu1,mu2) { cos(t)*mu1+sin(t)*mu2 }

#
# Logaritmo natural del kernel de la densidad f(ln r|theta).
llrt<-function(y,t,mu1,mu2) {
  2*y-0.5*exp(y)*(exp(y)-2*Dbd(t,mu1,mu2) )
}

140
## Valores iniciales para el algoritmo de Metropolis.
#
media0<-function(t,mu1,mu2) {
  log( ( Dbd(t,mu1,mu2) + ( (Dbd(t,mu1,mu2)^2) + 8 ) ^0.5 )/2 )
}

#
var0<-function(m0) {
  ( 2 + exp(2*m0) )^(-1)
}
#-----
## Algoritmo de Metrópolis - Hasting.

Metro<-function(f,t,mu1,mu2,tamuestra,nodeite)
{
  N<-tamuestra
  ite<-nodeite
  #
  m0<-media0(t,mu1,mu2)
  v0<-var0(m0)
  y0<-rnorm(N,m0,sqrt(v0))
  #
  for (i in 1:ite)
  {
    y1<-rnorm(N,m0,sqrt(v0))
    lfy1 <- f(y1,t,mu1,mu2)
    ldny1 <- log(dnorm(y1,m0,sqrt(v0)))
  141
    w1<-(lfy1-ldny1)
    lfy0 <- f(y0,t,mu1,mu2)
    ldny0 <- log(dnorm(y0,m0,sqrt(v0)))
    w0<-(lfy0 - ldny0)
    lalpha<-(w1-w0)
    u<-runif(N,0,1)
    aux<-ifelse(log(u)<=lalpha,y1,y0)
    y0<-aux
  }
  rdt<-exp(y0)
  drop(rdt)
}

```

```
}

#-----
rNPxy.1<-function(M,V){
  # Esta funcion simula una observacion (r,theta)
  # con vector de medias M y matriz de covarianza V.
  #
  library(MASS)
  xy<-mvrnorm(1,M,V)
  rr<-sqrt(sum(xy^2))
  ttheta<-(atan2(xy[2],xy[1]))%%(2*pi)
  xy.rt<-c(rr,ttheta)
  drop(xy.rt)
}
#-----
circ.stats<-function(data,flag) {
  # Esta funcion calcula la direccion media muestral (en grados)
  # flag: Si los datos estan en grados, ajustar flag=0.
  # Si los datos estan en radianes, ajustar flag=1.
  if (flag==0) data<-data*(pi/180)
  Cbarra<-mean(cos(data))
  Sbarra<-mean(sin(data))
  Rbarra<-sqrt(Cbarra^2 + Sbarra^2)
  auxmean<-atan(Sbarra/Cbarra)
  meanphi<-0*c(1:length(data))
  if (Cbarra<0) meanphi<-auxmean+pi else {
    if (auxmean<0) meanphi<-auxmean+(2*pi) else meanphi<-auxmean }
  theta.barra<-meanphi*180/pi
  drop(theta.barra)
}
#-----
# Matriz de diseño
X<-matrix(c(rep(1,n),x),n,2)
# X<-matrix(c(rep(1,n),x,x^2),n,3)
# Especificacion final para la "matriz" B.
XtX<-t(X)%*%X
Lstar1<-L1+XtX
Lstar2<-L2+XtX
Sigma1<-solve(Lstar1)
Sigma2<-solve(Lstar2)
L1m1<-L1%*%m1
L2m2<-L2%*%m2
# Matrices de diseño para cada componente.
p1<-7 # Dimensión del vector de efectos fijos.
p2<-7
```

```

q2<-1 # Dimensión del vector de efectos aleatorios.

XI<-array(0,c(N,n,p1))
XII<-array(0,c(N,n,p2))
for(i in 1:N){
  XI[, ,1]<-rep(1,n)
  XI[, ,2]<-rep(sh$sun[i],n)
  XI[, ,3]<-rep(sh$eye[i],n)
  XI[, ,4]<-rep(sh$w1[i],n)
  XI[, ,5]<-rep(sh$w2[i],n)
  XI[, ,6]<-rep(sh$w3[i],n)
  XI[, ,7]<-c(1:5)
  XII[, ,1]<-rep(1,n)
  XII[, ,2]<-rep(sh$sun[i],n)
  XII[, ,3]<-rep(sh$eye[i],n)
  XII[, ,4]<-rep(sh$w1[i],n)
  XII[, ,5]<-rep(sh$w2[i],n)
  XII[, ,6]<-rep(sh$w3[i],n)
  XII[, ,7]<-c(1:5) }
Z<-matrix(c(rep(1,n)))
#-----
# Parámetros de la especificación inicial.

A1<-matrix(0,p1,p1)
A2<-matrix(0,p2,p2)
v2<-q2
B2<-0.001
#-----
# Algunos objetos para la especificación de las distribuciones finales.

# En este caso, son iguales para cada componente.

XtX.I<-0.0
XtX.II<-0.0
for(i in 1:N){
  XtX.I<-t(XI[, ,])%*%XI[, ,]+ XtX.I # sum of { t(XI)*XI }
  XtX.II<-t(XII[, ,])%*%XII[, ,]+ XtX.II # sum of { t(XII)*XII }
}
ZtZ<-c(t(Z)%*%Z)
#-----
# Nota: Este código es pensando matricialmente.

#
etilde<-matrix(0,N,n)
b.aux<-matrix(0,N,q2)
bF<-matrix(0,N,q2)
#
D<-(ZtZ) + Omega
invD<-solve(D)

```

```

#
etilde<- Y - t( sapply( (1:N), function(w){X[w,,]%^%beta} ) )
#
bF<-t(invD%^%t(Z)%^%t(etilde))
b.aux<- as.matrix( sapply( (1:N), function(w){mvrnorm(1,bF[w,],invD)} ) )
# Nota: Si la dimension de b, cada {bi}, es mayor a 1, entonces de debe usar:
# b.aux<-t ( as.matrix( sapply( (1:N), function(w){mvrnorm(1,bF[w,],invD)} ) ) )
drop(b.aux)
}
b.f<-function(Omega,beta,Y){
# Muestreo de b (i.e. {bi} vectores, i=1,...,N)
# Lo siguiente es posible ya que en este caso Zi = Z para toda i.
#
etilde<-matrix(0,N,n)
b.aux<-c(1:N)*0.0
bF<-c(1:N)*0.0
D<-(ZtZ) + Omega
invD<-(1.0/D)
etilde<- Y - t( sapply( (1:N), function(w){XII[w,,]%^%beta} ) )
bF<-as.vector(t(invD*t(Z)%^%t(etilde)))
b.aux<- mvrnorm(1,bF,diag(N)*invD)
drop(b.aux)
}
#-----

```

## Apéndice C

Resultados obtenidos para el modelo elegido “eda\*coh + eda\*edu + res” mediante BRugs

	mean	sd	2.5%	50%	97.5%
eda2	-2,195	0,077	-2,348	-2,196	-2,040
eda3	-3,946	0,084	-4,108	-3,946	-3,785
eda3	0,011	0,076	-0,141	0,011	0,162
coh2	-0,168	0,083	-0,326	-0,168	0,000
coh3	-0,481	0,088	-0,654	-0,482	-0,311
edu2	-0,356	0,048	-0,449	-0,357	-0,261
res2	-0,298	0,108	-0,504	-0,297	-0,084
eda2coh2	-0,361	0,121	-0,608	-0,359	-0,129
eda2coh3	-0,399	0,124	-0,637	-0,401	-0,151
eda3coh2	-0,546	0,169	-0,879	-0,545	-0,217

eda3coh3	-0,171	0,125	-0,420	-0,172	0,071
eda2 edu2	-0,456	0,162	-0,786	-0,452	-0,151
eda3edu2	-0,587	0,057	-0,698	-0,587	-0,473
deviance	260,2	0,188	250,2	260,3	268,6

DIC info (using the rule, pD = Dbar-Dhat)

pD = 13,3 and DIC = 243,2

DIC is an estimate of expected predictive error (lower deviance is better).

## BIBLIOGRAFÍA

Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin. 1997  
“Bayesian Data Analysis”. Second Edition. Chapman and Hall.

Bernardo, J. & Smith, A. 1994. “Bayesian Theory”. First Edition. Wiley & Sons, New York.

Box, G. E. Tiao, G. C. 1992 “Bayesian Inference in Statistical”. First Edition,  
“BUGS Project”. <http://www.mrc-bsu.cam.ac.uk/bug>. (MRC Biostatistics Unit, Cambridge, UK).

Casella, G., and R.L.Berger. 1990. “Statistical Inference”. Duxbury Press.

Coa, R., y L.H. Ochoa. 2009. “Bolivia: Encuesta Nacional de Demografía y Salud. ENDSSA 2008”. Ministerio de Salud y Deportes, e Instituto Nacional de Estadística.

Celade, División de Población de la Cepal, Unicef, Unfpa. 2011. “Mortalidad en la niñez. Una base de datos de América Latina desde 1960”.

Cox, D.R., and D.V. Hinkley. 1974. “Theoretical Statistics”. Chapman and Hall.

Dean, C., and J.F. Lawless. June 1989. “Tests for Detecting Overdispersion in Poisson Regression Models”. American Statistical Association. Vol. 84. No. 406.

Lee, P. M. 2001 “Bayesian Statistics: An Introduction.” First Edition. Wiley.

McCullagh, P., and J.A. Nelder. 1989. "Generalized Linear Models". Second Edition. Chapman and Hall.

Pierce, D.A., and D.W. Schafer. December 1986. "Residuals in Generalized Linear Models". American Statistical Association. Vol. 81. No. 396.

Publicación de las Naciones Unidas (1983), Manual X. "Indirect Techniques for Demographic Estimation" (ST/ESA/SER. A/81), Nueva York. No. de venta E.83.XIII.2.

Robert C., and Casella G. 2004 "Monte Carlo Statistical methods". Second Edition. Springer.

XV Grupo consultivo 2012 "Revisión de la Estrategia Boliviana de Reducción de la Pobreza 2008 - 2012" BOLIVIA. Una alianza hacia las Metas del Milenio. Segunda edición 2012. UDAPE.