PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

INFERENCIA BAYESIANA EN EL MODELO DE REGRESIÓN BETA RECTANGULAR

Tesis para optar el grado de Magíster en Estadística que presenta

FRANCISCO GERMÁN CALDERÓN POZO

Dirigido por

DR. CRISTIAN LUIS BAYES RODRÍGUEZ

San Miguel, 2018

Dedicatoria

A mis padres Francisco y Yone por su permanente apoyo en todo momento de mi vida y por animarme a seguir estudiando.

A mi hermana Ingrid por compartir el gusto por las ciencias. Espero que nunca dejes de estudiar porque todos los días se aprende algo nuevo y en la carrera que has elegido, es necesario estar siempre actualizada.

A mis profesores y compañeros de la maestría en Estadística que me ayudaron a aprender un poco más de esta maravillosa ciencia.

Agradecimentos

En primer lugar quiero agradecer a mis padres Francisco y Yone, por los valores inculcados y por haberme brindado la oportunidad de tener una excelente educación en el transcurso de mi vida. Considero que ellos son un excelente ejemplo de vida a seguir.

Quiero expresar mi más profundo y sincero agradecimiento al Doctor Cristian Bayes, mi asesor de tesis, por la orientación, el seguimiento y la supervisión continúa. Gracias a él, pude incrementar mis conocimientos en los temas tratados en la presente tesis; asimismo, aumentaron mis habilidades en el manejo de *R*, *OpenBUGS*, *WinBUGS* y *JAGS*.

Agradezco a todos mis profesores de la Maestría en Estadística, quienes me enseñaron las herramientas para desarrollar este tema. En especial a los responsables del seminario de tesis: los doctores Cristian Bayes, Giancarlo Sal y Rosas, Jose Flores y Luis Valdivieso; quienes con sus preguntas, sugerencias y conocimientos aportaron a la mejora de mi investigación. También, a mis compañeros del seminario, quienes me ayudaron a resolver algunas dudas relacionadas a cálculos e implementación de los códigos en los *softwares* utilizados.

Debo agradecer también a mis buenos amigos Elio Peralta, Marylía Cruz y Juan Barzola, con quienes estudiaba para las prácticas, exámenes y me juntaba para realizar los trabajos grupales. Agradezco la paciencia y apoyo incondicional que siempre tienen conmigo.

Por último, pero no menos importante, le agradezco a Dios por haberme acompañado y guiado a lo largo de mi carrera, por brindarme una vida llena de experiencias, aprendizajes y felicidad.

ESCUELA DE
POSGRADO

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

RESUMEN DE LA TESIS

Francisco Germán Calderón Pozo
Maestría en Estadística
Inferencia Bayesiana en el Modelo de Regresión Beta Rectangular

.....

Se conoce que el modelo lineal normal no es apropiado para situaciones en la que la variable respuesta es una proporción que solo toma valores en un rango limitado $(0, 1)$, pues, se pueden obtener valores ajustados para la variable de interés que exceden sus límites inferior y superior.

Ante dicha situación, una propuesta es utilizar la distribución beta ya que es bastante flexible para modelar proporciones. Este modelo de regresión, sin embargo, puede ser influenciado por la presencia de valores atípicos o extremos. Debido a ello, se ha propuesto en la literatura, un modelo de mayor robustez llamado modelo de regresión beta rectangular, el cual permite una mayor incidencia de tales valores.

El objetivo general de la tesis es estudiar las propiedades, estimar y aplicar a un conjunto de datos reales el modelo de regresión beta rectangular desde el punto de vista de la estadística bayesiana.

Para cumplir con el objetivo planteado, se estudian las características y propiedades de las distribuciones beta y beta rectangular. Luego, se desarrolla el análisis bayesiano del modelo de regresión beta rectangular considerando las distribuciones a priori y a posteriori, los criterios de selección de modelos y simulaciones de Montecarlo vía cadenas de Markov. También, se realizan estudios de simulación para demostrar que el nuevo modelo es más robusto que el modelo de regresión beta. Adicionalmente, se presenta una aplicación para mostrar la utilidad del modelo de regresión beta rectangular.

Palabras-clave: Regresión beta rectangular, regresión beta, inferencia Bayesiana, valores extremos.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En diversas disciplinas como la Economía, Ingeniería y Psicología, por citar algunas; se investiga la influencia de ciertas covariables en una variable continua que admite únicamente valores en el intervalo $(0, 1)$, tales como, proporciones, ratios, tasas, etc. Por ejemplo, la tasa de mortalidad de una cierta región puede ser influenciada por la endemicidad de malaria o el porcentaje de casos de tuberculosis ocurridos en cierto período de tiempo y región puede ser influenciado por la proporción de pobres, cantidad de personas vacunadas, número de campañas de prevención, etc.

En estos casos, como mencionan Ferrari y Cribari-Neto (2004), los modelos de regresión usuales pueden no ser apropiados para situaciones en la que la variable respuesta solo toma valores en un rango limitado $(0, 1)$; debido a que se pueden obtener valores ajustados para la variable de interés que exceden sus límites inferior y superior.

Una clase de distribución que permite modelar variables continuas limitadas al intervalo $(0, 1)$ es la distribución beta, la cual es bastante flexible para modelar este tipo de datos debido a que su función de densidad puede tomar diversas formas dependiendo del valor de los dos parámetros que caracterizan a esta distribución.

Sin embargo, el modelo de regresión beta planteado por Ferrari y Cribari-Neto (2004) puede ser influenciado por la presencia de valores extremos, motivo por el cual se han propuesto alternativas como el modelo de regresión beta rectangular propuesto por Bayes, Bazán y García (2012). Este nuevo modelo permite una mayor probabilidad en la ocurrencia de eventos extremos considerando que la variable respuesta sigue la distribución beta rectangular planteada por Hahn (2008) la cual es una mezcla de una distribución beta con una distribución uniforme. Esta distribución podría verse como un caso particular de modelo mixto beta finito propuesto por Bouguila et al. (2006). Es bien conocido que las distribuciones mixtas son más robustas ante valores atípicos o extremos (se refiere a valores más grandes o influyentes) debido a que incluyen una componente de distribución extra; más aún, la variabilidad se explica mejor y la estimación del «verdadero» parámetro de la media es menos afectada, como indica Markatou (2000).

1.2. Objetivos

El objetivo general de la tesis es estudiar propiedades, estimar y aplicar a un conjunto de datos reales el modelo de regresión beta rectangular desde el punto de vista de la estadística bayesiana. De manera específica se busca:

- Revisar la literatura acerca de las diferentes propuestas de modelos de regresión para proporciones.
- Estudiar a profundidad la distribución beta rectangular.
- Estudiar las características y propiedades del modelo de regresión beta rectangular desde la perspectiva bayesiana.
- Implementar métodos de inferencia bayesiana considerando simulación MCMC.
- Realizar estudios de simulación para evaluar el desempeño del modelo de regresión beta rectangular en diferentes escenarios.
- Aplicar el modelo a un conjunto de datos reales.

1.3. Organización del trabajo

En el Capítulo 2, se estudiarán las características y propiedades de la distribución beta. En el Capítulo 3, se estudiará a profundidad la distribución beta rectangular. En el Capítulo 4, se desarrollará el análisis bayesiano del modelo de regresión beta rectangular considerando las distribuciones a priori y a posteriori, así como los criterios de comparación para estimadores y los criterios de selección de modelos. En el Capítulo 5 se muestra un estudio de simulación considerando los modelos de regresión beta y beta rectangular con la finalidad de demostrar que este último es más robusto. En el Capítulo 6 se trabaja una aplicación en la cual se analiza como se ve afectada la tasa de analfabetismo de los distritos del departamento de La Libertad por la condición de pobreza en el año 2007.

Finalmente, en el Capítulo 7 se discuten algunas conclusiones obtenidas en este trabajo. También, se brindan algunas recomendaciones para realizar investigaciones futuras.

En el apéndice A se muestran los programas desarrollados en el *software* R para realizar las simulaciones de datos considerando el paquete *rjags*. En el apéndice B se presentan los códigos en R donde se utiliza el paquete *rjags*. Asimismo, en el apéndice C se muestra el conjunto de datos utilizado en el caso aplicativo. Finalmente, en el apéndice D se presentan los gráficos de valores simulados y de no existencia de autocorrelación de los modelos obtenidos en la aplicación.

Capítulo 2

Distribución beta

La distribución beta permite modelar variables continuas que toman valores en el intervalo $(0, 1)$, lo que la hace muy apropiada para modelar proporciones. Por ejemplo: la tasa de analfabetismo de un país, la PEA ocupada según actividad económica, la proporción de defectuosos en un determinado lote, etc. Por otro lado, en la inferencia bayesiana es muy utilizada como distribución a priori cuando las observaciones tienen una distribución binomial.

En este capítulo se describirá a la distribución beta, se estudiarán sus propiedades y se presentará una parametrización alternativa propuesta por Ferrari y Cribari-Neto (2004).

2.1. Función de densidad de probabilidad

La función de densidad de probabilidad de una variable aleatoria Y que sigue una distribución beta es dada por

$$g_Y(y \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1, \quad \alpha > 0, \quad \beta > 0. \tag{2.1}$$

Una de las principales ventajas de esta distribución es el ajuste a una gran variedad de distribuciones empíricas, pues adopta formas muy diversas dependiendo de cuáles sean los valores de los parámetros de forma α y β, mediante los que viene definida la distribución.

En la Figura 2.1 se muestran las gráficas de la distribución beta para algunos valores de α y β. Se puede apreciar en **(1)** que con un valor de $\alpha = 1$ y $\beta > 1$, la curva tiene forma de J invertida. En **(2)** se visualiza que cuando α y β son iguales a 0.5, la distribución beta tiene forma de U. En **(3)** se visualiza que cuando los valores de α y β son iguales a 6 se tiene una curva simétrica. En **(4)** se muestra que cuando el valor de $\alpha = 6$ y $\beta = 3$, la curva tiene asimetría negativa o de cola a la izquierda.

2.2. Propiedades de la distribución beta

La media y la varianza de la distribución beta son expresadas por

$$E(Y) = \frac{\alpha}{\alpha+\beta} \quad \text{y} \quad Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{2.2}$$

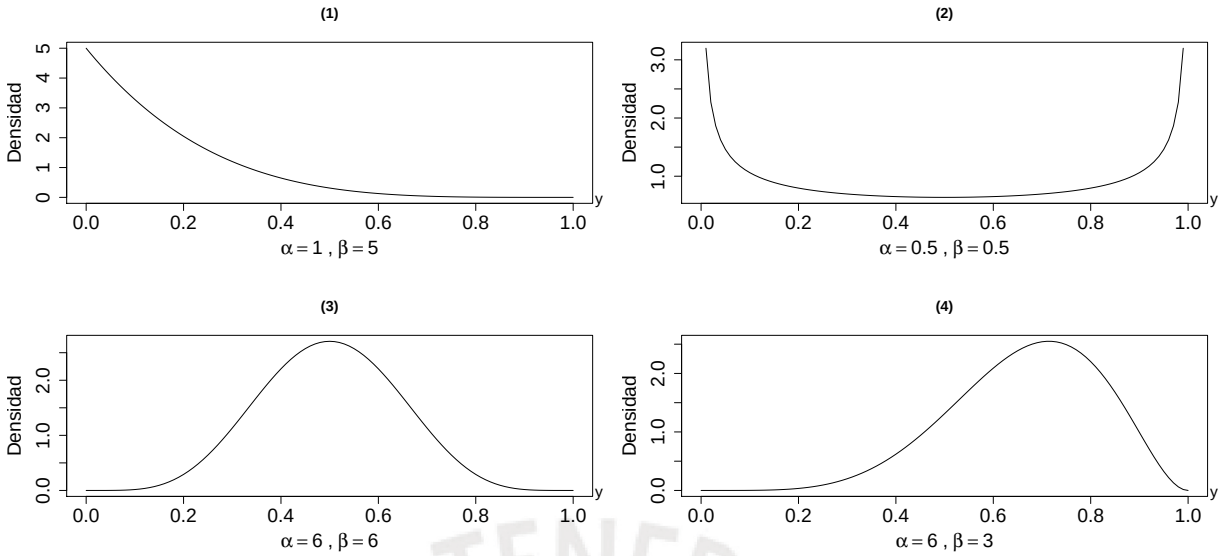


Figura 2.1: Función de densidad de la distribución beta para diferentes valores de los parámetros.

La distribución beta tiene una moda solo si ambos de sus parámetros son mayores que 1. La expresión de la moda es la siguiente:

$$Moda = \frac{\alpha - 1}{\alpha + \beta - 2}, \quad \text{si} \quad \alpha, \beta > 1 \tag{2.3}$$

La desviación absoluta según Krishnamoorthy (2006) se define como una medida de la variabilidad de los posibles valores de la variable aleatoria Y, y se expresa como $E(|Y - \mu|)$. Para esta distribución está dada por:

$$Desviación\ absoluta = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{2\alpha^{\alpha}\beta^{\beta}}{(\alpha + \beta)^{\alpha+\beta+1}}. \tag{2.4}$$

El coeficiente de simetría es la división del tercer momento respecto a la media entre la varianza elevada a $\frac{3}{2}$, para la distribución beta se expresa de la siguiente manera:

$$Simetria = \frac{2(\beta - \alpha)\sqrt{\alpha + \beta + 1}}{(\alpha + \beta + 2)\sqrt{\alpha\beta}}. \tag{2.5}$$

Del mismo modo, el coeficiente de kurtosis es el cuarto momento respecto a la media entre la varianza elevada al cuadrado y se expresa de la siguiente forma:

$$Kurtosis = \frac{3(\alpha + \beta + 1)[2(\alpha + \beta)^2 + \alpha\beta(\alpha + \beta - 6)]}{\alpha\beta(\alpha + \beta + 2)(\alpha + \beta + 3)}. \tag{2.6}$$

Dos casos importantes de mencionar son el de Beta(1,1) que es equivalente a la distribución uniforme en el intervalo unitario y el caso límite cuando ambos α y $\beta \to 0$, que corresponde

a la distribución Bernoulli(p), donde p es determinado por el ratio de convergencia relativa de α y β a 0 como se menciona en Smithson y Verkuilen (2006).

2.3. Parametrización alternativa de la distribución beta

Muchos profesionales comúnmente utilizan modelos de regresión para analizar datos que están relacionados con otras variables. El modelo de regresión lineal es el más usado; sin embargo no es apropiado para situaciones en las que la variable respuesta está restringida al intervalo $(0, 1)$, debido a que se pueden obtener valores ajustados para la variable de interés que exceden los límites superiores e inferiores.

Debido a ello, en la literatura se han propuesto modelos de regresión basados en la distribución beta, porque tiene una alta flexibilidad para modelar proporciones debido a que su función de densidad puede, como vimos, tomar diferentes formas al variar sus dos parámetros.

Uno de los más famosos modelos de regresión fue propuesto por Ferrari y Cribari-Neto (2004). El modelo propuesto por estos autores permite la interpretación de los parámetros en términos de la respuesta en su escala original (variable respuesta sin transformar). Ellos consideraron una reparametrización del modelo donde $\mu = \alpha/(\alpha + \beta)$ y $\phi = \alpha + \beta$, siguiendo (2.2) se obtiene que

$$E(Y) = \mu \quad \text{y} \quad Var(Y) = \frac{V(\mu)}{1 + \phi} \tag{2.7}$$

donde $V(\mu) = \mu(1 - \mu)$; de tal manera que μ es la media de la variable respuesta y ϕ puede ser interpretado como un parámetro de precisión en el sentido que para valores fijos de μ, al aumentar el valor de ϕ, la varianza de Y disminuye. En esta nueva parametrización la densidad de la distribución beta puede ser escrita como

$$b_Y(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \tag{2.8}$$

donde $0 < \mu < 1$ y $\phi > 0$.

Tal como se mencionó en la Sección 2.1, la distribución beta está asociada a los parámetros de forma α y β; sin embargo, tal como se menciona en esta sección, esta parametrización no es adecuada para llevar a cabo un análisis de regresión; por ello, se realizó una parametrización alternativa que asocia la distribución beta a sus parámetros de media y precisión.

En la Figura 2.2 se pueden observar diferentes densidades beta para los valores correspondientes de (μ, ϕ). La curva es simétrica cuando $\mu = 0.5$ y asimétrica cuando $\mu \neq 0.5$. También, puede tomar la forma de J y J invertida como se visualiza en el panel central. En el tercer panel se puede observar que la distribución beta reparametrizada tiene la forma de U cuando $\mu = 0.5$ y $\phi = 1$; por otro lado, cuando $\mu = 0.5$ y $\phi = 2$, se obtiene una distribución uniforme. La dispersión de la distribución para el parámetro fijo μ, disminuye cuando el valor de ϕ aumenta. Otro aspecto importante es que a medida que aumenta el valor de ϕ, las colas se

vuelven más pesadas.

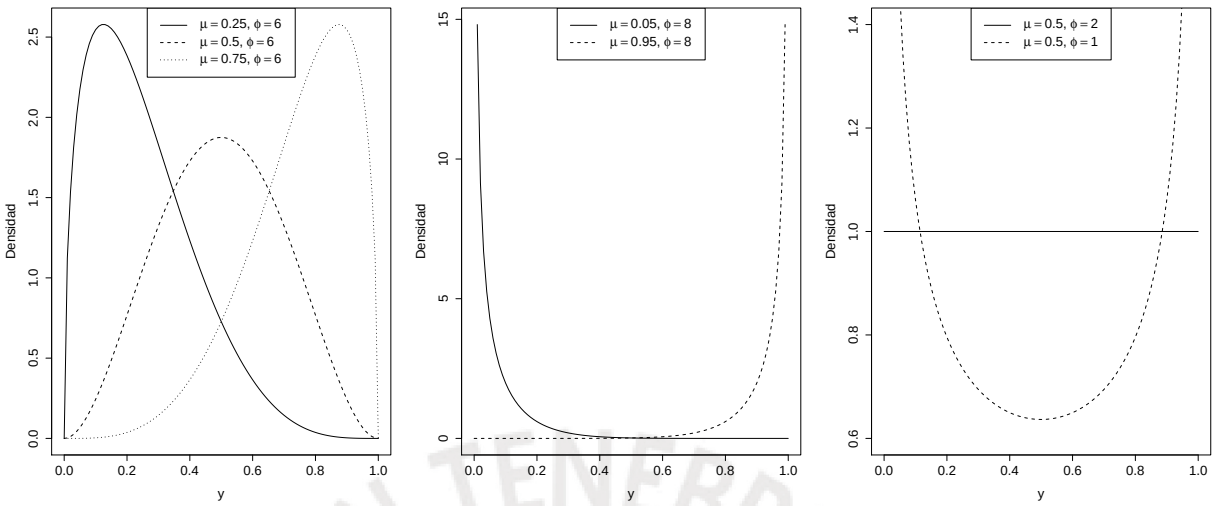


Figura 2.2: Función de densidad de la distribución beta reparametrizada para diferentes valores de los parámetros μ y ϕ.

Considerando esta nueva parametrización y la ecuación (2.3), la moda de la distribución beta reparametrizada se puede escribir de la siguiente manera:

$$Moda = \frac{\mu\phi - 1}{\phi - 2}, \quad \text{si} \quad \mu\phi > 1 \quad \text{y} \quad (1-\mu)\phi > 1 \tag{2.9}$$

La desviación absoluta puede ser re-escrita a partir de la ecuación (2.4) de la siguiente manera:

$$Desviación\ absoluta = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} \frac{2(\mu\phi)^{\mu\phi}((1-\mu)\phi)^{(1-\mu)\phi}}{\phi^{\phi+1}}. \tag{2.10}$$

Tomando en cuenta que $V(\mu) = \mu(1-\mu)$, el coeficiente de simetría para la distribución beta reparametrizada puede ser escrito a partir de (2.5) como:

$$Simetria = \frac{2(1-\mu)(\phi+1)^{0.5}}{(\phi+2)V(\mu)^{0.5}}. \tag{2.11}$$

El coeficiente de kurtosis es re-escrito a partir de la ecuación (2.6) como:

$$Kurtosis = \frac{3(\phi+1)[2+V(\mu)(\phi-6)]}{V(\mu)(\phi+2)(\phi+3)}. \tag{2.12}$$

Capítulo 3

Distribución beta rectangular

La distribución beta rectangular fue propuesta por Hahn (2008), debido a la necesidad de contar con una distribución con rango acotado y que presente colas pesadas. De modo que le permitiera modelar en forma más flexible el tiempo de duración de una actividad dentro de un proyecto. En particular, Hahn (2008) aplicó esta distribución como parte de la herramienta de gestión de proyectos *PERT (Program Evaluation and Review Technique)*

Esta nueva distribución es una mezcla de una distribución beta con una distribución uniforme; que como muestran Bayes et al. (2012) permite realizar inferencia más robusta ante la presencia de valores atípicos.

En este capítulo se mostrará la función de densidad de probabilidad, sus propiedades y una parametrización alternativa de la distribución beta rectangular propuesta por Bayes et al. (2012).

3.1. Función de densidad de probabilidad

La función de densidad de probabilidad de una variable aleatoria Y que sigue una distribución beta rectangular es dada por:

$$f_Y(y \mid \mu, \phi, \theta) = (1-\theta)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} + \theta. \tag{3.1}$$

Esta nueva distribución utiliza los parámetros μ y ϕ que provienen de la parametrización alternativa propuesta en la ecuación (2.8); e incorpora un parámetro de mixtura $0 \leq \theta \leq 1$. Se puede notar que la distribución uniforme se obtiene cuando $\theta = 1$ y la distribución beta se encuentra cuando $\theta = 0$. Se usará la notación $Y \sim BR(\mu, \phi, \theta)$ para hacer referencia a la distribución beta rectangular de una variable aleatoria Y con parámetros μ, ϕ y θ.

La distribución propuesta puede ser escrita siguiéndose la notación empleada por Bayes et al. (2012) de la siguiente manera:

$$f_Y(y \mid \mu, \phi, \theta) = \theta + (1-\theta) b_Y(y \mid \mu, \phi), \tag{3.2}$$

donde $b_Y(y \mid \mu, \phi)$ está definida en (2.8).

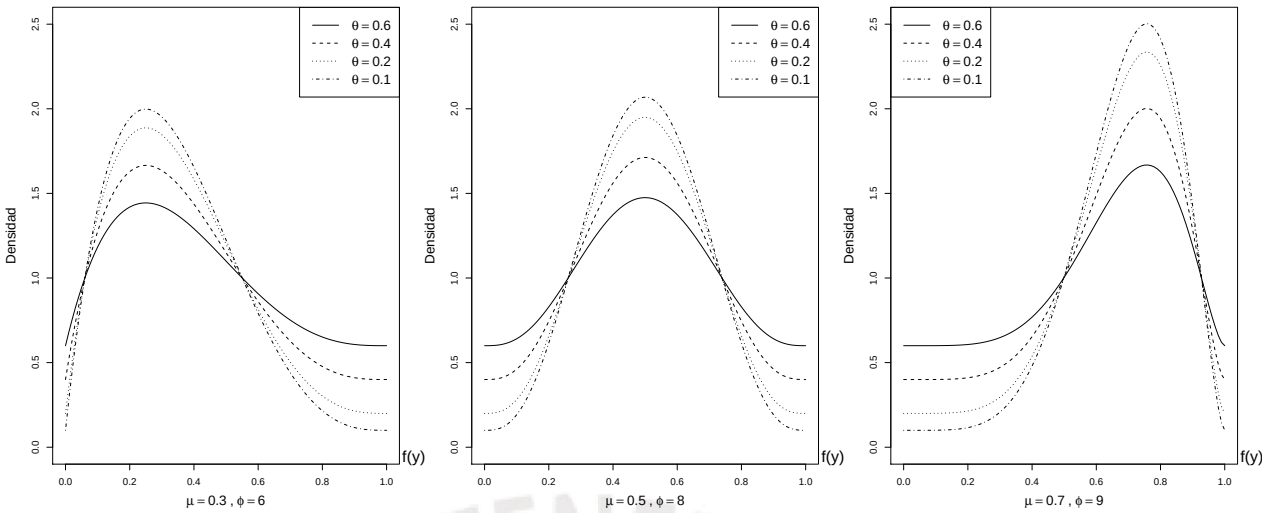


Figura 3.1: Función de densidad de la distribución beta rectangular para diferentes valores de los parámetros μ, ϕ y θ

En la Figura 3.1 se pueden observar las diversas formas adoptadas por la función de densidad, dependiendo del valor de los parámetros que caracterizan a esta distribución. En el gráfico del lado izquierdo, se visualiza que con un valor de $\mu = 0.3$, la curva tiene asimetría positiva (o a la derecha); en el gráfico central, la curva es simétrica cuando el valor de μ es 0.5 y en el gráfico del lado derecho, la curva tiene asimetría negativa (o a la izquierda) cuando $\mu = 0.7$. En todos los gráficos se visualiza que a medida que el valor de θ aumenta, las colas se vuelven más pesadas.

3.2. Propiedades de la distribución beta rectangular

La media y varianza de la distribución beta rectangular son expresadas por

$$E(Y) = \frac{\theta}{2} + (1-\theta)\mu, \tag{3.3}$$

$$Var(Y) = \frac{V(\mu)}{1+\phi}(1-\theta)(1+\theta(1+\phi)) + \frac{\theta}{12}(4-3\theta). \tag{3.4}$$

Se debe notar que cuando $\theta = 0$, las expresiones dadas en (3.3) y (3.4) coinciden con las expresiones de la media y la varianza de la distribución beta(ver Bayes et al., 2012):

$$E(Y) = \mu \quad \text{y} \quad Var(Y) = \frac{V(\mu)}{1+\phi}. \tag{3.5}$$

Por otro lado, cuando $\theta = 1$, las expresiones dadas en (3.3) y (3.4) coinciden con las expresiones de la media y la varianza de la distribución uniforme:

$$E(Y) = \frac{1}{2} \quad \text{y} \quad Var(Y) = \frac{1}{12}. \tag{3.6}$$

En la distribución beta rectangular, cuando $0 \leq \theta \leq 1$, la moda está dada por:

$$Moda = \frac{\mu\phi - 1}{\phi - 2}, \quad \text{si} \quad \mu\phi > 1 \quad \text{y} \quad (1-\mu)\phi > 1 \tag{3.7}$$

que coincide con la moda de la distribución beta mostrada en (2.9).

3.3. Parametrización alternativa de la distribución beta rectangular

Como se mencionó anteriormente, Bayes et al. (2012) notan que la media de la distribución beta rectangular dada en la ecuación (3.3) es una función de los parámetros θ y μ. Al igual que con la distribución beta, se considera a la media de esta distribución, γ, como un nuevo parámetro, de tal manera que:

$$E(Y) = \frac{\theta}{2} + (1-\theta)\mu = \gamma,$$

luego, se puede escribir μ en términos de γ y θ para obtener lo siguiente:

$$\mu = \frac{\gamma - \frac{\theta}{2}}{1-\theta}. \tag{3.8}$$

Como se conoce que μ se encuentra entre 0 y 1; se impone la siguiente desigualdad:

$$\frac{\theta}{2} < \gamma < 1 - \frac{\theta}{2}, \tag{3.9}$$

Se puede observar en la inecuación (3.9) que el espacio paramétrico de γ depende de θ. Entonces, se plantean las siguientes condiciones para el parámetro θ:

$$\begin{cases} 0 < \theta < 2\gamma \quad \text{si} \quad \gamma < \frac{1}{2}, \\ 0 < \theta < 2(1-\gamma) \quad \text{si} \quad \gamma > \frac{1}{2} \end{cases} \tag{3.10}$$

Al utilizar la parametrización empleada por Bayes et al. (2012), se tiene que para que $\gamma \in [0,1]$, θ debe cumplir la siguiente condición:

$$0 < \theta < 1 - |2\gamma - 1|. \tag{3.11}$$

Con la finalidad de obtener una estructura más apropiada para modelar la media de la distribución beta rectangular, Bayes et al. (2012) proponen una nueva reparametrización que sustituye los parámetros θ y μ por los parámetros

$$\gamma = \frac{\theta}{2} + (1-\theta)\mu \quad \text{y} \quad \alpha = \frac{\theta}{1-|2\gamma-1|}, \tag{3.12}$$

Considerando la nueva parametrización dada en (3.12), la media y la varianza de la distribución beta rectangular pueden ser escritas como:

$$E(Y) = \gamma, \tag{3.13}$$

$$Var(Y) = \frac{\gamma(1-\gamma)}{(1+\phi)(1-\theta)}(1+\theta(1+\phi)) - \frac{\theta}{2}(1-\frac{\theta}{2})\frac{1+\theta(1+\phi)}{(1+\phi)(1-\theta)} + \frac{\theta}{12}(4-3\theta), \tag{3.14}$$

donde $\theta = \alpha(1-2|\gamma-1|)$. Nótese que en este caso, los espacios paramétricos de γ y α son independientes con $\gamma \in (0,1)$ y $\alpha \in (0,1)$.

Bajo esta parametrización, la función de densidad de probabilidad de la distribución beta rectangular, considerando la notación $Y \sim BRr(\gamma, \phi, \alpha)$, está dada por (Bayes et al., 2012):

$$h_Y(y|\gamma,\phi,\alpha) = \alpha(1-|2\gamma-1|) + (1-\alpha(1-|2\gamma-1|))b\left(y|\frac{\gamma-0.5\alpha(1-|2\gamma-1|)}{1-\alpha(1-|2\gamma-1|)}, \phi\right) \tag{3.15}$$

Esta función permite tener mayor robustez en la estimación de los parámetros porque permite incluir los valores extremos que se puedan presentar en un conjunto de datos (ver Chia, 2012).

En la Figura 3.2 se muestra la función de densidad de la distribución beta rectangular reparametrizada para distintos valores del parámetro γ, especificados en cada panel. En el panel (1) se muestra el gráfico de la función beta rectangular simétrica, que se da cuando $\gamma = 0.5$; en el panel (2) se muestra que esta función puede tomar formas asimétricas, como en este caso que la media es igual a $\gamma = 0.2$. En el panel (3) se muestra la función en forma de J invertida, que se da cuando por ejemplo $\gamma = 0.05$ y en el panel (4) se muestra la función en forma de J, que se da cuando $\gamma = 0.95$.

Al igual que con la parametrización original, mientras el valor de α sea más cercano a 0, la función toma la forma de la distribución beta; por otro lado, cuando el parámetro se acerca a 1, la función es más plana, es decir se parece más a la distribución uniforme (ver la línea continua punteada).

En la Figura 3.3 se muestra la función de densidad de la distribución beta rectangular reparametrizada para $\gamma = 0.5$ y diferentes valores de α, ϕ en cada uno de los paneles; siendo evidente que la distribución tiene colas más pesadas a medida que aumenta el valor de α. El parámetro ϕ puede ser entendido como un parámetro de precisión; por ello, la dispersión de la distribución va disminuyendo a medida que dicho parámetro aumenta. El panel (1) muestra la mayor dispersión del ejemplo debido a que $\phi = 5$ y la menor dispersión proviene

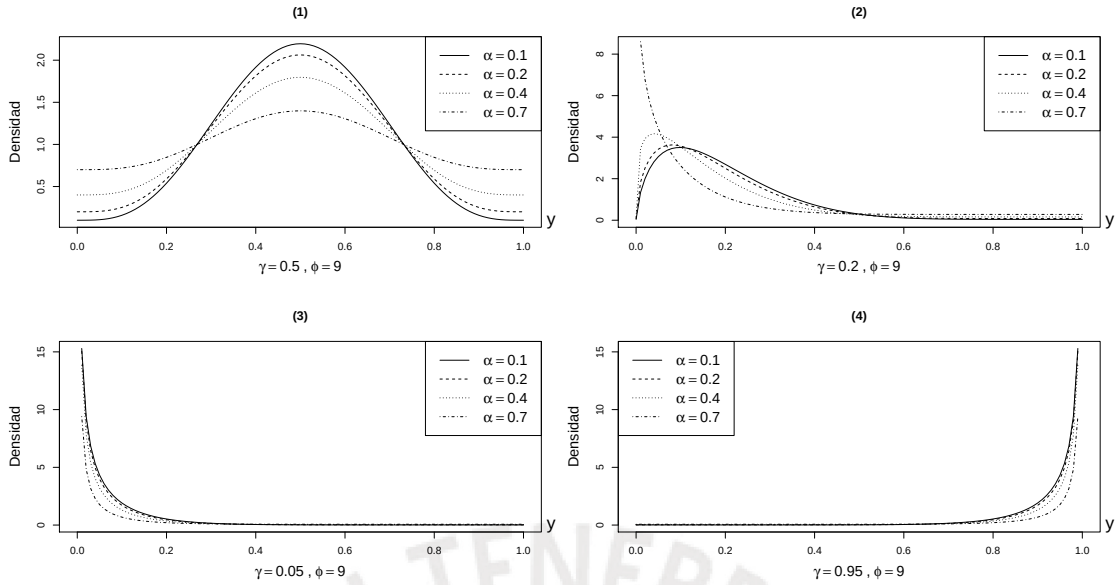


Figura 3.2: Función de densidad de la distribución beta rectangular bajo la parametrización dada en 3.15 para diferentes valores de los parámetros.

del panel (4) con un valor de $\phi = 150$.

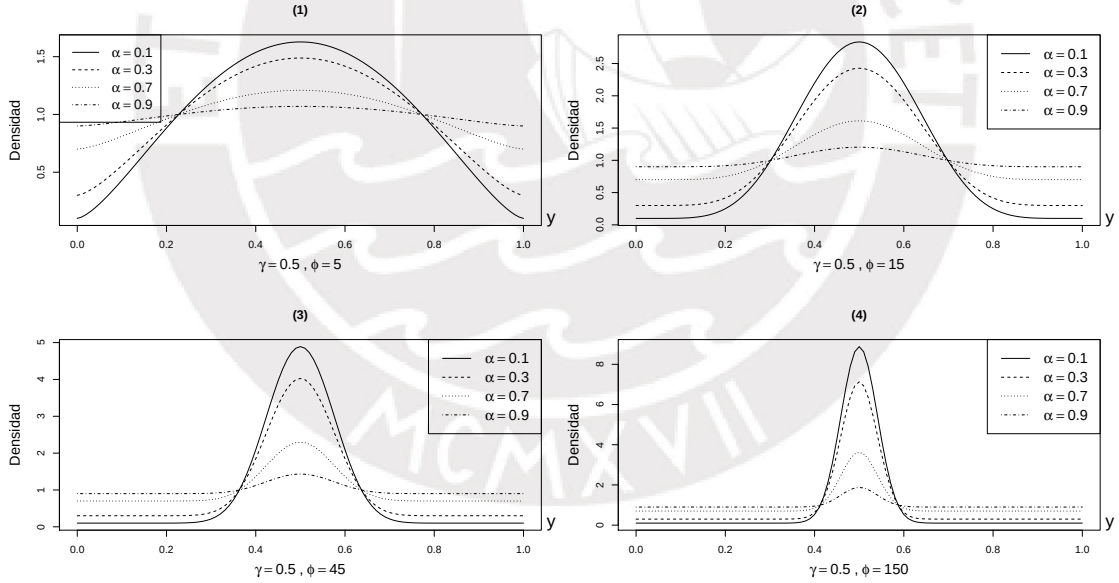


Figura 3.3: Función de densidad de la distribución beta rectangular bajo la parametrización dada en 3.15 para valores de $\gamma = 0.5$, $\phi = 5$ (panel 1), $\phi = 15$ (panel 2), $\phi = 45$ (panel 3), $\phi = 150$ (panel 4).

Es necesario resaltar que a diferencia de la parametrización original, α no es un parámetro de mixtura sino, como señalan Bayes et al. (2012), un parámetro de forma que está asociado con la amplitud de la cola. Mientras que ϕ es un parámetro que controla la precisión de la distribución, de tal manera que mientras mayor sea su valor, se observa menor dispersión.

3.4. Espacio paramétrico de las distribuciones beta rectangular y beta rectangular reparametrizada

La parametrización propuesta por Hahn (2008) no es adecuada para aquellos casos en los que se debe hacer análisis de regresión, por lo que Bayes et al. (2012) propusieron la reparametrización de la distribución beta rectangular descrita en la Sección 3.3. También, se observó que el espacio paramétrico presentado por Hahn (2008) estaba restringido por el espacio dado en la desigualdad (3.11). Los parámetros γ y α en cambio se mueven de manera independiente en el cuadrado unitario con $\gamma \in (0, 1)$ y $\alpha \in (0, 1)$.

En la Figura 3.4 se muestra el espacio paramétrico propuesto por Bayes et al. (2012) y estudiado por Chia (2012). En el panel izquierdo se muestra el espacio para los parámetros θ y γ; mientras que el espacio para α y γ se muestra en el panel derecho.

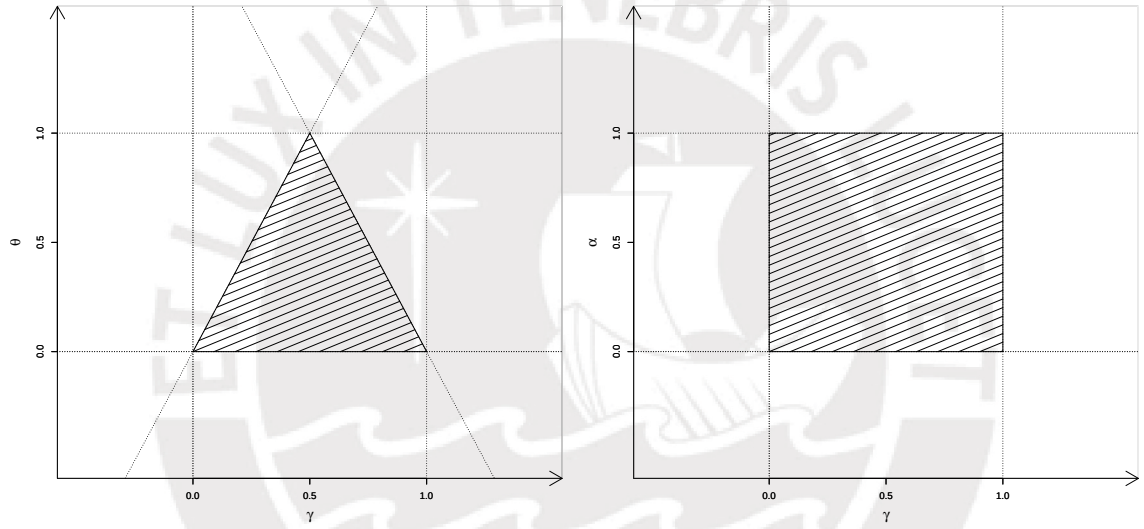


Figura 3.4: Espacio paramétrico para la distribuciones Beta Rectangular (γ,θ) y Beta Rectangular Reparametrizada(γ,α). Fuente: Chia (2012).

En el panel izquierdo se puede apreciar que el espacio en el cual θ está definido no permite que γ tome ciertos valores, lo cual no es tan beneficioso para realizar la regresión. En cambio, al realizar la reparametrización planteada en la expresión (3.12) se obtiene un espacio en el que γ tiene mayor alcance tal como se visualiza en el panel derecho.

Capítulo 4

Inferencia bayesiana en el modelo de regresión beta rectangular

Los modelos de regresión tratan de estimar o predecir el valor de una variable dependiente en función de valores conocidos de variables explicativas. En el presente capítulo se muestra el modelo de regresión beta rectangular desde la perspectiva bayesiana. También, se tratan aspectos como el ajuste del modelo, las distribuciones a priori y la elección del modelo basado en los criterios de comparación.

4.1. Modelo de regresión beta rectangular

La distribución beta es considerada flexible, sin embargo, como observaron Hahn (2008) y Bayes et al. (2012), esta distribución no considera los eventos extremos.

Bayes et al. (2012) reafirmaron y demostraron que el modelo de regresión beta está fuertemente influenciado en la estimación de los parámetros de regresión cuando hay observaciones atípicas en la variable respuesta; por ello, sugirieron un nuevo modelo, llamado beta rectangular, que sea robusto frente a este tipo de observaciones.

El modelo de regresión para el vector de respuestas observadas $\mathbf{Y} = (Y_1, ..., Y_n)^T$ que siguen una distribución beta rectangular está dado por:

$$\begin{array}{rcl} Y_i & \sim & BRr(\gamma_i, \phi_i, \alpha) \\ F_1^{-1}(\gamma_i) & = & \mathbf{x}_i^T \boldsymbol{\beta} \\ F_2^{-1}(\phi_i) & = & -\mathbf{w}_i^T \boldsymbol{\delta} \end{array} \tag{4.1}$$

donde $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ y $\boldsymbol{\delta} = (\delta_1, ..., \delta_l)^T$ son vectores de parámetros de regresión, $\mathbf{x}_i = (x_{i1}, ..., x_{ik})^T$ y $\mathbf{w}_i = (w_{i1}, ..., w_{il})^T$ son valores de k y l covariables. Asimismo, $F_1^{-1}(\cdot)$ y $F_2^{-1}(\cdot)$ son funciones reales estrictamente monótonas y doblemente diferenciables en $(0, 1)$ con dominios en $\mathbb{R}$ para el primer caso y $\mathbb{R}^+$ para el segundo caso.

En general $F_1(\cdot)$ puede ser cualquier función de distribución acumulada que corresponda a una distribución continua tal que su función inversa, llamada de enlace, relaciona el parámetro de la media γ_i con las covariables $\mathbf{x}_i$. Las funciones de enlace para $F_1^{-1}(\cdot)$ pueden ser logit, probit, entre otras. $F_2^{-1}(\cdot)$ es una función de enlace que relaciona el parámetro de precisión

ϕ_i con las covariables $\mathbf{w}_i$; en este caso se utiliza el enlace $log(\phi_i) = -\mathbf{w}_i^T\boldsymbol{\delta}$, el signo negativo fue indicado por Smithson y Verkuilen (2006) para que la interpretación del coeficiente δ sea más fácil. Dado que ϕ es un parámetro de precisión, el signo positivo de δ indica una menor varianza, lo cual puede llevar a confusión. El signo negativo permite modelar la dispersión en vez de la precisión que es una acción más común.

4.1.1. Función de verosimilitud

Considerando la reparametrización propuesta por Bayes et al. (2012) definida en (3.15) se tiene que la función de verosimilitud es dada por:

$$L(\boldsymbol{\eta}|Y) = \prod_{i=1}^{n} h_Y(y_i|\gamma_i, \phi_i, \alpha)$$

donde

$$h_Y(y_i|\gamma_i,\phi_i,\alpha) = \alpha(1-|2\gamma_i-1|)+(1-\alpha(1-|2\gamma_i-1|))b\left(y_i|\frac{\gamma_i-0.5\alpha(1-|2\gamma_i-1|)}{1-\alpha(1-|2\gamma_i-1|)},\phi_i\right) \quad (4.2)$$

$\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T, \alpha)$, γ_i y ϕ_i están definidas en (4.1) con $\theta_i = \alpha(1-2|\gamma_i - 1/2|)$ y $\mu_i = \dfrac{\gamma_i - \theta_i/2}{1-\theta_i}$.

Se puede apreciar en (3.15) que cuando $\alpha = 0$ y ϕ_i es constante, se obtiene el modelo de regresión beta propuesto por Ferrari y Cribari-Neto (2004) que no está en la familia exponencial y por tanto no es un Modelo Lineal Generalizado (MLG). También, se obtiene el modelo de regresión beta con dispersión variable introducido por Paolino (2001), Smithson y Verkuilen (2006) y Simas et al. (2010).

4.1.2. Función de verosimilitud aumentada

Dado que el modelo beta rectangular es una mixtura finita entre una distribución beta y una distribución uniforme, este admite la siguiente reparametrización jerarquica:

$$\begin{aligned} Y_i|Z_i = 1 &\sim Uniforme(0,1) \\ Y_i|Z_i = 0 &\sim Beta(\mu_i, \phi_i) \\ Z_i &\sim Bernoulli(\theta_i) \end{aligned}$$

donde Z_i es una variable auxiliar o latente. Este tipo de variables se usan para modelar las características no observadas de los sujetos que determinan las respectivas probabilidades de éxito. Asimismo, θ_i y μ_i fueron definidos en (4.1) y (4.2).

Luego, si $Z = (Z_1, Z_2, ..., Z_n)$, entonces la función de verosimilitud aumentada del modelo de regresión beta rectangular puede escribirse como:

$$\begin{aligned} L(\boldsymbol{\eta}, Z|Y) &= \prod_{i=1}^{n} f_Y(y_i|z_i) f_Y(z_i) \\ &= \prod_{i=1}^{n} b_Y(y_i|\mu_i, \phi)^{1-z_i} 1^{z_i} \theta_i^{z_i} (1-\theta_i)^{1-z_i} \\ &= \prod_{i=1}^{n} b_Y(y_i|\mu_i, \phi)^{1-z_i} \theta_i^{z_i} (1-\theta_i)^{1-z_i}, \end{aligned} \tag{4.3}$$

donde $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T, \alpha)$.

4.2. Distribución a priori

En el presente trabajo de investigación se considerará el supuesto de que no se cuenta con información previa acerca de los parámetros (β, δ, α).

Para $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T, \alpha^T)$ se usarán entonces las siguientes distribuciones a *priori* (ver Bayes et al., 2012):

$$\begin{aligned} \beta &\sim Normal(\mathrm{a}, \mathbf{B}) \\ \delta &\sim Normal(\mathrm{c}, \mathbf{D}) \\ \alpha &\sim Uniforme(0, 1) \end{aligned} \tag{4.4}$$

donde a y c son vectores de dimensiones k y l respectivamente; mientras que $\mathbf{B}$ y $\mathbf{D}$ son matrices $k \times k$ y $l \times l$ respectivamente. Asimismo, se asume que los elementos del vector de parámetros son independientes (ver Bayes et al., 2012); por lo tanto la distribución a *priori* es dada por:

$$p(\boldsymbol{\eta}) = p(\boldsymbol{\beta}) p(\boldsymbol{\delta}) p(\alpha).$$

De aquí en adelante, se usará la notación $N(\mu, \sigma^2)$ para hacer referencia a la distribución Normal y la notación $U(a, b)$ para hacer referencia a la distribución uniforme.

4.3. Distribución a posteriori

La distribución a posteriori es proporcional a la multiplicación de la priori propuesta en (4.4) por la función de verosimilitud y se define como:

$$p(\boldsymbol{\eta}|\mathbf{Y}) \propto L(\boldsymbol{\eta}|\mathbf{Y}) p(\boldsymbol{\eta})$$

Si se considera que $\beta \sim N_k(a, B)$, $\delta \sim N_l(c, D)$ y $\alpha \sim U(0, 1)$; la forma de la distribución a posteriori es la siguiente:

$$\begin{aligned} p(\boldsymbol{\eta}|\mathbf{Y}) \propto & \prod_{i=1}^{n} \alpha(1-|2F_1(\mathbf{x}_i^T\boldsymbol{\beta})-1|) + (1-\alpha|(1-|2F_1(\mathbf{x}_i^T\boldsymbol{\beta})-1|)) \\ & \times b\left(y_i|\frac{F_1(\mathbf{x}_i^T\boldsymbol{\beta})-0.5\alpha(1-|2F_1(\mathbf{x}_i^T\boldsymbol{\beta})-1|)}{1-\alpha(1-|2F_1(\mathbf{x}_i^T\boldsymbol{\beta})-1|)}, F_2(\mathbf{w}_i^T\boldsymbol{\delta})\right) \\ & \times \phi_k(\boldsymbol{\beta}|a,\mathbf{B}) \times \phi_l(\boldsymbol{\delta}|c,\mathbf{D}) \times 1 \end{aligned} \tag{4.5}$$

donde $\phi_k(\cdot)$ y $\phi_l(\cdot)$ expresan la función de densidad de una distribución normal multivariada de orden k y l respectivamente.

Asimismo, la función de distribución a posteriori aumentada se expresa como:

$$p(\boldsymbol{\eta}, Z|\mathbf{Y}) \propto L(\boldsymbol{\eta}, Z|\mathbf{Y})p(\boldsymbol{\eta})$$

Para este caso, se considera también $\beta \sim N_k(a, B)$, $\delta \sim N_l(c, D)$ y $\alpha \sim U(0,1)$; de tal manera que la distribución a posteriori aumentada es la siguiente:

$$\begin{aligned} p(\boldsymbol{\eta}, Z|\mathbf{Y}) \propto & \prod_{i=1}^{n} b(y_i|\mu_i,\phi)^{1-z_i}\theta_i^{z_i}(1-\theta_i)^{1-z_i} \\ & \times \phi_k(\boldsymbol{\beta}|a,\mathbf{B}) \times \phi_l(\boldsymbol{\delta}|c,\mathbf{D}) \times 1 \end{aligned} \tag{4.6}$$

donde $\mu_i = \dfrac{\gamma_i - \theta_i/2}{1-\theta_i}$ y $\phi_i = e^{-\mathbf{w}_i^T\boldsymbol{\delta}}$.

Se usará esta distribución a posteriori porque es más sencilla de implementar computacionalmente y permite trabajar con data en diferentes períodos.

4.4. Criterios de comparación para selección de modelos

Actualmente existen diferentes metodologías para comparar modelos bajo inferencia bayesiana, entre las que destacan el *EAIC (Esperado del Criterio de Información de Akaike)* y el *EBIC (Esperado del Criterio de Información Bayesiano)*. Ambos fueron propuestos por Brooks (2002). También, se utiliza el *DIC* (Criterio de Información del Desvío) que fue propuesto por Spiegelhalter et al. (2002).

Se tiene que $D(\boldsymbol{\eta}) = -2logL(\boldsymbol{\eta}|y)$ representa el desvío, donde $L(\boldsymbol{\eta}|y)$ es la función de verosimilitud expresada en (4.2). Entonces, Spiegelhalter et al. (2002) define el DIC como:

$$DIC = E_{\boldsymbol{\eta}|y}[D(\boldsymbol{\eta})] + \rho_D$$

donde $E_{\boldsymbol{\eta}|y}[D(\boldsymbol{\eta})]$ representa la media a posteriori del desvío y ρ_D es el número efectivo de parámetros definido como:

$$\rho_D = E_{\boldsymbol{\eta}|y}[D(\boldsymbol{\eta})] - D[E_{\boldsymbol{\eta}|y}(\boldsymbol{\eta})],$$

donde $D[E_{\boldsymbol{\eta}|y}(\boldsymbol{\eta})]$ representa el desvío evaluado en la media a posteriori.

Mientras que Brooks (2002) define el $EAIC$ y $EBIC$ de la siguiente manera:

$$EAIC = E_{\boldsymbol{\eta}|y}[D(\boldsymbol{\eta})] + 2p$$
$$EBIC = E_{\boldsymbol{\eta}|y}[D(\boldsymbol{\eta})] + plog(n)$$

siendo p el número de parámetros en el modelo y n el número total de observaciones.

Luego, $E_{\boldsymbol{\eta}|y}[D(\boldsymbol{\eta})]$, $D[E_{\boldsymbol{\eta}|y}(\boldsymbol{\eta})]$ y ρ_D pueden ser estimadas usando los resultados de la simulación MCMC por:

$$\overline{D} = \frac{1}{B}\sum_{b=1}^{B} D(\boldsymbol{\eta}^b), \quad \widehat{D} = D(\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{\eta}^b) \text{ y } \widehat{\rho_D} = \overline{D} - \widehat{D},$$

respectivamente, donde B representa el número de iteraciones y $D(\boldsymbol{\eta}^b) = -2logL(\boldsymbol{\eta}^b|Y)$ es el valor de la desviación en la iteración b. Finalmente, los criterios $EAIC$, $EBIC$ y DIC son estimados por:

$$\widehat{EAIC} = \overline{D} + 2p, \quad \widehat{EBIC} = \overline{D} + plog(n) \text{ y } \widehat{DIC} = \overline{D} + \widehat{\rho_D} = 2\overline{D} - \widehat{D}.$$

Menores valores de DIC, $EBIC$ y $EAIC$ implican un mejor ajuste del modelo.

Capítulo 5

Estudio de simulación

En el presente capítulo se muestran los resultados de un estudio de simulación considerando los modelos de regresión beta y beta rectangular. El objetivo del estudio fue observar el comportamiento de ambos modelos frente a la presencia de datos atípicos; para ello, se compararon el sesgo y error cuadrático medio de los parámetros estimados y se obtuvo el porcentaje de casos en los que el modelo de regresión beta rectangular obtuvo un menor valor estimado de DIC. En el presente estudio se consideran 30 escenarios a partir de la combinación de distintos tamaños de muestra y porcentajes de valores extremos que se utilizan para crear diferentes patrones de perturbación. Para ello, se usan los *software* libres R en su versión 3.2.2 y $JAGS$ en su versión 3.4.0 similar a Alencar (2016). Los códigos computacionales se encuentran en el apéndice A.

5.1. Modelos en estudio

Similar a Bayes et al. (2012) y Alencar (2016) para el presente estudio de simulación se considerarán modelos de regresión con precisión constante.

5.1.1. Modelo de regresión beta

En primer lugar, se considera un modelo de regresión beta con las siguientes características:

$$\begin{aligned} Y_i &\sim Beta(\mu_i, \phi) \\ logit(\mu_i) &= \beta_1 + \beta_2 x_i, \end{aligned}$$

donde i=1,2,...,n. Las distribuciones a priori que se consideran para cada parámetro son: $\beta_1 \sim N(0, 100^2)$, $\beta_2 \sim N(0, 100^2)$ y $\phi \sim Gamma(0.01, 0.01)$.

5.1.2. Modelo de regresión beta rectangular

Luego, se considera un modelo de regresión beta rectangular que cumpla con lo siguiente:

$$\begin{aligned} Y_i &\sim BR(\gamma_i, \phi, \alpha) \\ logit(\gamma_i) &= \beta_1 + \beta_2 x_i, \end{aligned}$$

donde i=1,2,...,n. Las distribuciones a priori que se consideran para cada parámetro son: $\beta_1 \sim N(0, 100^2)$, $\beta_2 \sim N(0, 100^2)$, $\alpha \sim U(0, 1)$ y $\phi \sim Gamma(0.01, 0.01)$.

5.2. Criterios de comparación para estimadores

Existen diferentes criterios para evaluar el desempeño de un estimador. A continuación, se presentan tres (ver Gentle, 2002, cap.1):

1. Sesgo: El cual es definido como

$$sesgo(T) = E(T) - \theta.$$

El sesgo de un estimador es una medida de cuanto error se comete, en promedio, cuando se utiliza T para estimar el parámetro θ. Un estimador es insesgado si $E(T) = \theta$. Para determinar $E(T)$ se debe conocer la distribución de la estadística T. Cuando la distribución de T es desconocida se puede utilizar algún método de simulación para estimar el sesgo.

2. Error cuadrático medio: Este criterio se define como

$$ECM(T) = E[(T - \theta)^2].$$

El ECM se puede escribir también en función de la varianza y el sesgo de T

$$ECM(T) = Var(T) + [sesgo(T)]^2.$$

Cuando la varianza y el sesgo al cuadrado son pequeños, el ECM será pequeño. Si T es insesgado, entonces $ECM(T) = Var(T)$.

3. Error estándar: Es la desviación estándar de la distribución muestral de un estadístico y se define como

$$EE(T) = \sqrt{Var(T)} = \sigma_T.$$

5.3. Simulación de datos

En este estudio se establece una covariable X, tal que $X \sim U(-4, 4)$, lo cual se tomará en cuenta para la generación de un modelo de regresión beta dado por:

$$Y_i \sim Beta(\mu_i, \phi)$$
$$logit(\mu_i) = \beta_1 + \beta_2 x_i,$$

donde i=1,2,...,n.

Para este análisis, se considera $\beta_1 = 0.6$, $\beta_2 = 1$ y $\phi = 40$, tres tamaños de muestra

$n \in \{100, 150, 200\}$ y porcentajes de casos con valores extremos $r \in \{0\,\%, 2\,\%, 4\,\%, 6\,\%\}$. La combinación de valores de n y r produce $3 \times 4 = 12$ escenarios.

Se desea estudiar el efecto que producen los valores atípicos en la estimación de los parámetros de localización y el parámetro de precisión de los modelos beta y beta rectangular; para ello, se seleccionan $r \times n/100$ datos simulados. Luego, se sustituyen estos valores de y_i por $y_i^* = y_i \pm \Delta$, donde Δ es el incremento o decremento de los valores de y. En este trabajo se toman en cuenta 3 patrones de perturbación similar a Alencar (2016) y Bayes et al. (2012) :

(I) Un incremento de Δ unidades de los valores de y correspondiente a los $r \times n/100$ valores más bajos de x.

(II) Una disminución de Δ unidades de los valores de y correspondiente a los $r \times n/100$ valores más altos de x.

(III) Un incremento o disminución de Δ unidades de los valores de y correspondiente a los $r \times n/100$ valores más altos y más bajos de x respectivamente.

En primer lugar se consideran 3 escenarios sin data perturbada y luego se obtienen 27 escenarios con la data contaminada para cada modelo. En la Figura 5.1 se muestra la data obtenida cuando $n = 200$, $r = 2\,\%$ para la data sin contaminar y con datos atípicos. Para el patrón de perturbación (I) se utilizó un Δ de 0.9, para el patrón de perturbación (II) se utilizó un Δ de 0.80 y para el patrón de perturbación (III) se utilizó un Δ incremental de 0.90 y un Δ de disminución de 0.80. En los demás escenarios se utilizó el mismo procedimiento para obtener los datos con valores atípicos.

Para cada escenario y para cada conjunto de datos, se estimaron los modelos beta y beta rectangular tal como fueron explicados en la Sección 5.1. Esta estimación se realiza utilizando el programa *JAGS* que implementa un algoritmo de Gibbs. Este necesita una fase inicial de muestreo durante la cual los muestreadores adaptan su comportamiento para maximizar su eficiencia; en este caso el período de adaptación considerado fue de 5,000 iteraciones. El período de *burn-in* considerado fue de 25,000 para 100,000 valores de la cadena y para reducir la autocorrelación se tomaron solamente saltos de 10 en 10 (*thin* =10).

El sesgo, el error cuadrático medio y el error estándar fueron calculados para cada modelo considerando las réplicas en cada escenario; también, se obtuvo el porcentaje de casos en el cual el modelo de regresión beta rectangular fue seleccionado en vez del modelo de regresión beta. Esto se define en términos del porcentaje de veces en el cual el modelo de regresión beta rectangular obtiene un menor valor estimado de DIC. Cabe resaltar que en los escenarios donde no existe data perturbada, el modelo de regresión beta obtiene un menor valor estimado de DIC en un mayor porcentaje de veces.

En las Figuras 5.2 y 5.3 se muestra para cada patrón que a medida que r aumenta, el sesgo para β_1 y β_2 es mayor; además, el modelo de regresión beta rectangular muestra menores valores de sesgo que se acercan a cero. Asimismo, en las Figuras 5.4 y 5.5 se muestra que a medida que el porcentaje de datos atípicos aumenta, el ECM aumenta; sin embargo, el

modelo de regresión beta rectangular presenta menores valores de ECM. En general, existe una mejora en la precisión (el sesgo y el error cuadrático medio decrecen) para las estimaciones de los parámetros cuando se emplea un modelo de regresión beta rectangular en vez de un modelo de regresión beta, esta observación se sustenta en que el modelo de regresión beta rectangular obtiene un menor valor estimado de DIC en un gran porcentaje de los casos; tal como se observa en la última columna de la Tabla 5.1.

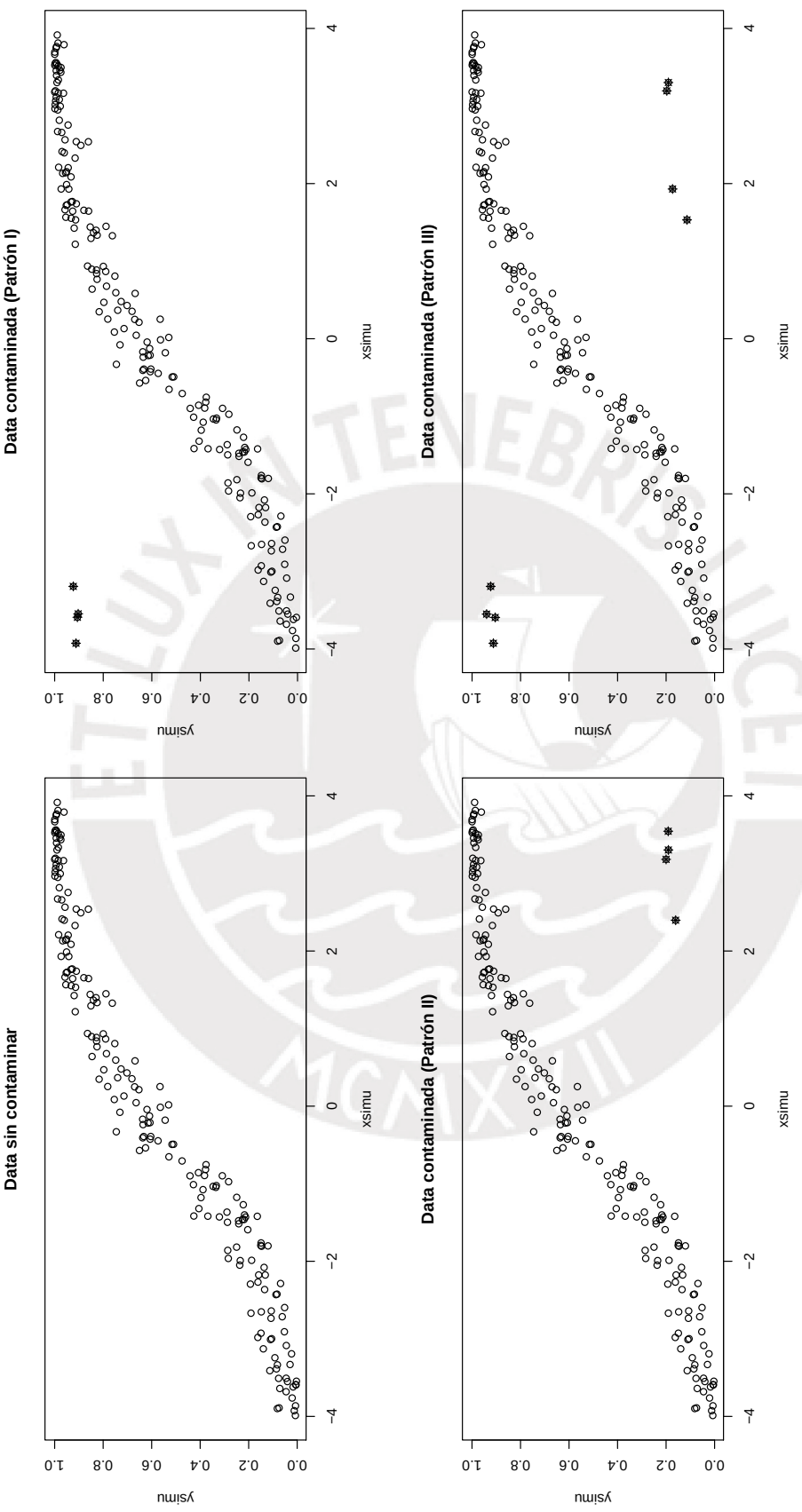
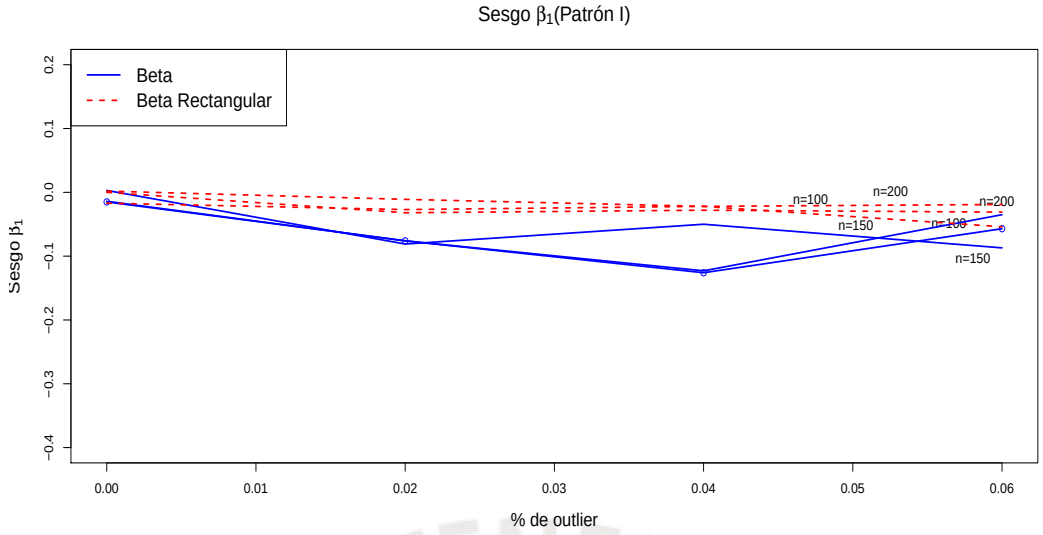


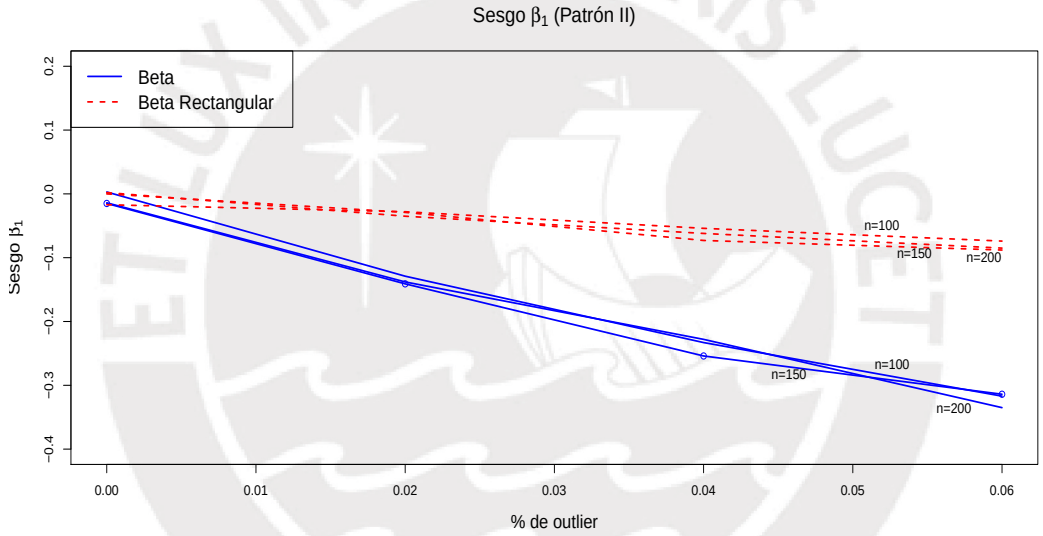
Figura 5.1: Generación de data sin contaminar y con valores extremos (representados con un *): (I) incremento de Δ unidades de los valores y correspondiente a valores más bajos de x, (II) disminución de Δ unidades de los valores y correspondiente a valores más altos de x, (III) incremento o disminución de Δ unidades de los valores y correspondiente a valores más altos y más bajos de x respectivamente.

	Escenarios			Beta						Beta rectangular						
	Patrones	% de atípicos	n	Sesgo β_1	Sesgo β_2	ECM β_1	ECM β_2	EE β_1	EE β_2	Sesgo β_1	Sesgo β_2	ECM β_1	ECM β_2	EE β_1	EE β_2	DIC
1	-	0	100	-0.015	0.000	0.002	0.001	0.044	0.033	0.000	-0.016	0.002	0.001	0.043	0.029	0.25
2	-	0	150	0.003	-0.005	0.001	0.000	0.039	0.020	0.002	-0.016	0.001	0.001	0.031	0.025	0.35
3	-	0	200	-0.014	-0.007	0.001	0.000	0.028	0.014	-0.017	-0.007	0.001	0.000	0.034	0.015	0.30
4	Patrón I	2 %	100	-0.076	-0.235	0.007	0.056	0.033	0.022	-0.032	-0.048	0.003	0.003	0.047	0.020	0.65
5	Patrón I	2 %	150	-0.081	-0.231	0.007	0.054	0.032	0.015	-0.011	-0.051	0.002	0.003	0.047	0.029	0.70
6	Patrón I	2 %	200	-0.076	-0.225	0.006	0.051	0.020	0.013	-0.027	-0.052	0.002	0.003	0.035	0.016	0.60
7	Patrón I	4 %	100	-0.126	-0.367	0.017	0.134	0.033	0.011	-0.028	-0.078	0.002	0.007	0.039	0.025	0.85
8	Patrón I	4 %	150	-0.050	-0.385	0.003	0.148	0.028	0.013	-0.022	-0.080	0.001	0.007	0.028	0.024	0.65
9	Patrón I	4 %	200	-0.123	-0.352	0.016	0.124	0.030	0.013	-0.022	-0.088	0.001	0.008	0.030	0.019	0.95
10	Patrón I	6 %	100	-0.057	-0.469	0.005	0.221	0.040	0.017	-0.031	-0.117	0.003	0.014	0.050	0.024	1.00
11	Patrón I	6 %	150	-0.087	-0.463	0.008	0.214	0.029	0.011	-0.054	-0.137	0.005	0.019	0.044	0.025	1.00
12	Patrón I	6 %	200	-0.035	-0.479	0.002	0.229	0.021	0.006	-0.019	-0.121	0.001	0.015	0.032	0.028	1.00
13	Patrón II	2 %	100	-0.141	-0.171	0.021	0.030	0.037	0.024	-0.029	-0.048	0.002	0.003	0.038	0.026	0.85
14	Patrón II	2 %	150	-0.129	-0.168	0.017	0.028	0.025	0.018	-0.035	-0.045	0.002	0.002	0.027	0.018	0.65
15	Patrón II	2 %	200	-0.138	-0.168	0.020	0.028	0.029	0.013	-0.028	-0.043	0.002	0.002	0.042	0.021	0.85
16	Patrón II	4 %	100	-0.254	-0.279	0.066	0.078	0.041	0.015	-0.073	-0.092	0.007	0.009	0.038	0.023	1.00
17	Patrón II	4 %	150	-0.233	-0.274	0.055	0.076	0.021	0.016	-0.062	-0.087	0.005	0.008	0.034	0.026	0.65
18	Patrón II	4 %	200	-0.228	-0.274	0.052	0.075	0.023	0.017	-0.054	-0.073	0.003	0.006	0.024	0.016	1.00
19	Patrón II	6 %	100	-0.314	-0.361	0.100	0.130	0.041	0.019	-0.088	-0.121	0.009	0.015	0.038	0.025	1.00
20	Patrón II	6 %	150	-0.317	-0.363	0.101	0.132	0.021	0.013	-0.085	-0.110	0.008	0.013	0.031	0.022	0.85
21	Patrón II	6 %	200	-0.335	-0.365	0.113	0.134	0.024	0.010	-0.074	-0.111	0.007	0.013	0.038	0.024	1.00
22	Patrón III	2 %	100	-0.106	-0.198	0.013	0.040	0.044	0.017	-0.037	-0.043	0.004	0.003	0.049	0.030	1.00
23	Patrón III	2 %	150	-0.088	-0.153	0.009	0.024	0.032	0.013	-0.018	-0.036	0.002	0.002	0.043	0.022	1.00
24	Patrón III	2 %	200	-0.110	-0.200	0.013	0.040	0.033	0.016	-0.034	-0.049	0.002	0.003	0.030	0.021	0.90
25	Patrón III	4 %	100	-0.158	-0.326	0.026	0.107	0.038	0.017	-0.022	-0.088	0.002	0.008	0.034	0.027	1.00
26	Patrón III	4 %	150	-0.164	-0.322	0.028	0.104	0.037	0.017	-0.038	-0.091	0.002	0.009	0.030	0.028	0.65
27	Patrón III	4 %	200	-0.191	-0.322	0.037	0.104	0.028	0.016	-0.044	-0.079	0.003	0.007	0.030	0.031	1.00
28	Patrón III	6 %	100	-0.159	-0.422	0.026	0.179	0.028	0.018	-0.062	-0.117	0.007	0.014	0.054	0.024	1.00
29	Patrón III	6 %	150	-0.207	-0.378	0.044	0.143	0.030	0.015	-0.049	-0.099	0.004	0.010	0.038	0.022	0.90
30	Patrón III	6 %	200	-0.200	-0.418	0.040	0.175	0.024	0.009	-0.061	-0.130	0.004	0.017	0.026	0.022	1.00

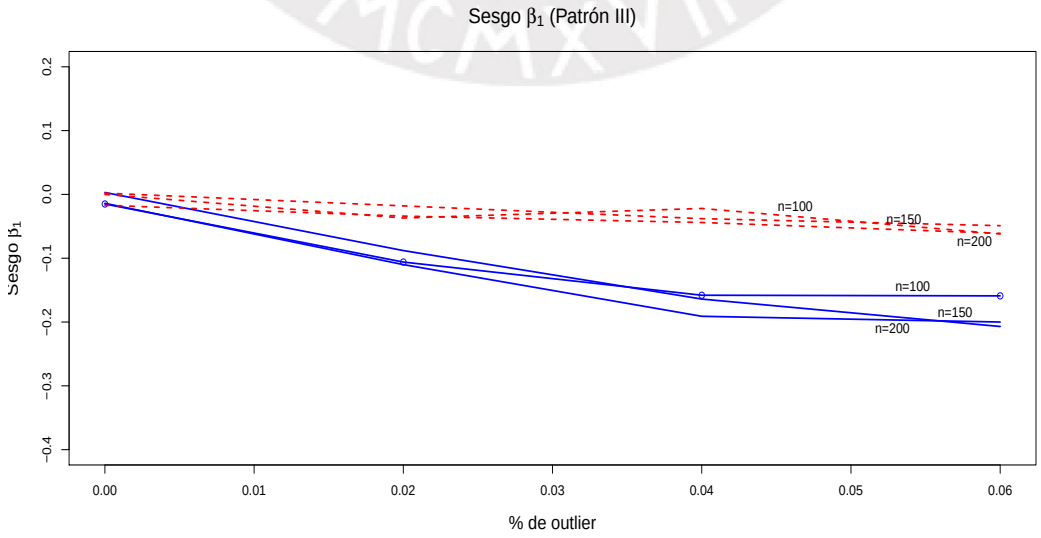
Tabla 5.1: Comparación de sesgo, error cuadrático medio, error estándar, DIC y porcentaje de selección entre el modelo de regresión beta rectangular y el modelo de regresión beta considerando diferentes escenarios, data contaminada (con $\phi = 40$, 3 % de valores atípicos y 3 tamaños de muestra) y 100 simulaciones bajo cada escenario.



(a) Sesgo de β_1 siguiendo el patrón I para diferentes tamaños de n

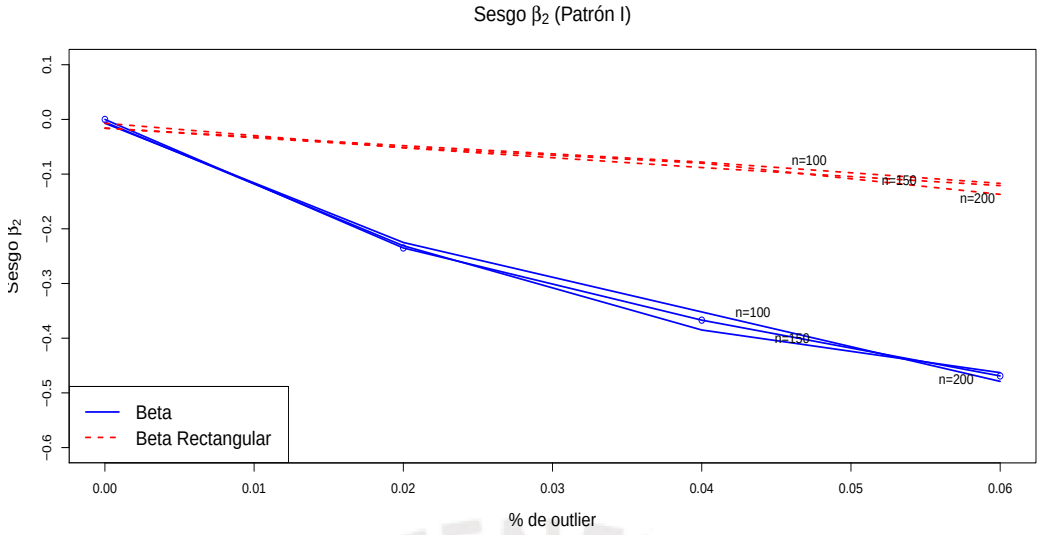


(b) Sesgo de β_1 siguiendo el patrón II para diferentes tamaños de n

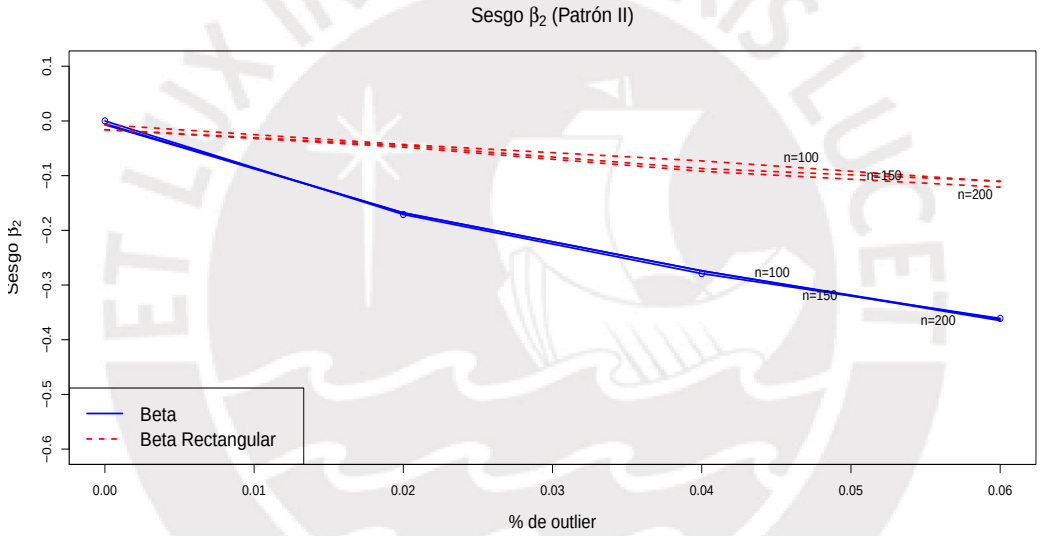


(c) Sesgo de β_1 siguiendo el patrón III para diferentes tamaños de n

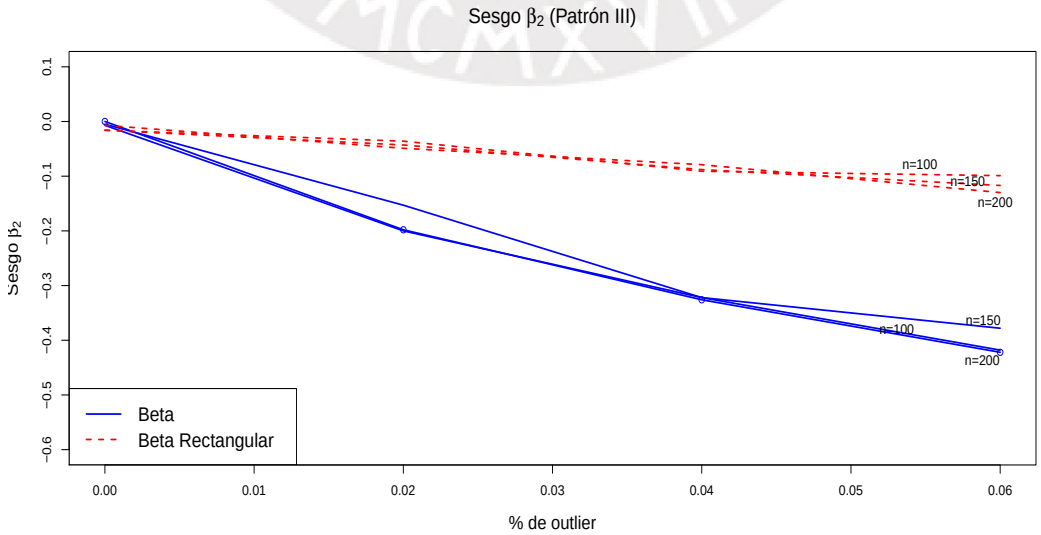
Figura 5.2: Gráficos de comparación de sesgo para β_1 obtenidos para los modelos de regresión beta y beta rectangular considerando diferentes tamaños de n y porcentajes de valores extremos.



(a) Sesgo de β_2 siguiendo el patrón I para diferentes tamaños de n

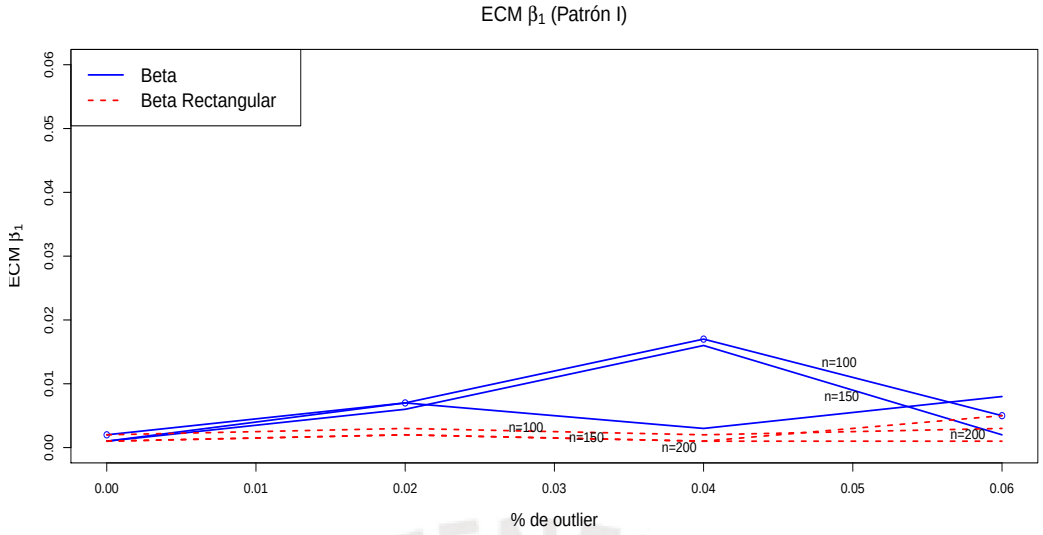


(b) Sesgo de β_2 siguiendo el patrón II para diferentes tamaños de n

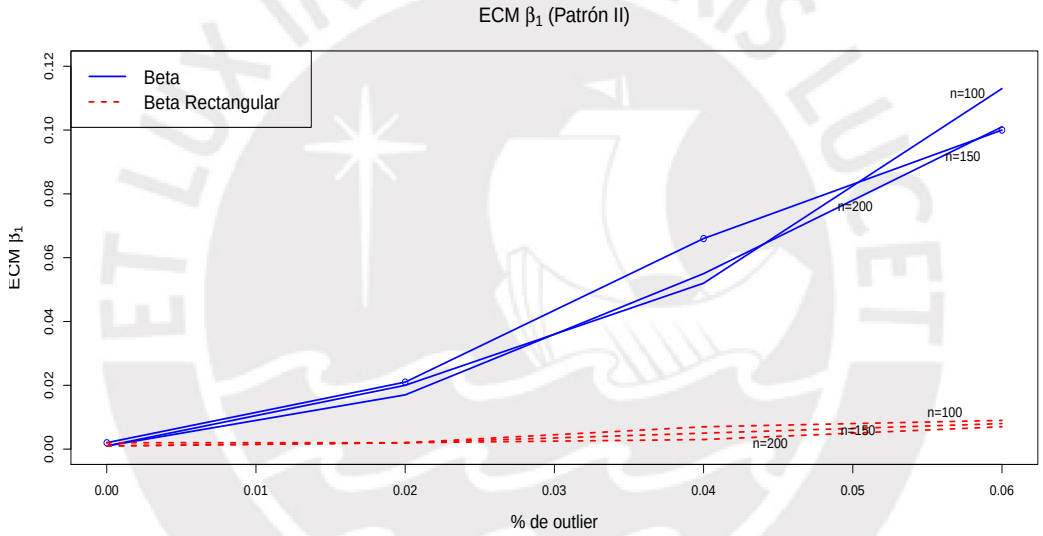


(c) Sesgo de β_2 siguiendo el patrón III para diferentes tamaños de n

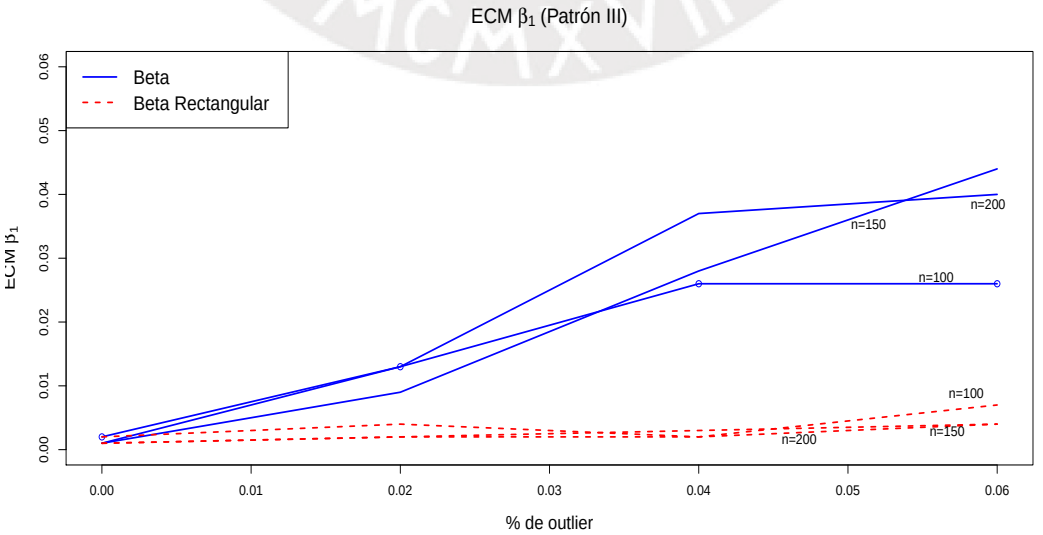
Figura 5.3: Gráficos de comparación de sesgo para β_2 obtenidos para los modelos de regresión beta y beta rectangular considerando diferentes tamaños de n y porcentajes de valores extremos.



(a) ECM de β_1 siguiendo el patrón I para diferentes tamaños de n

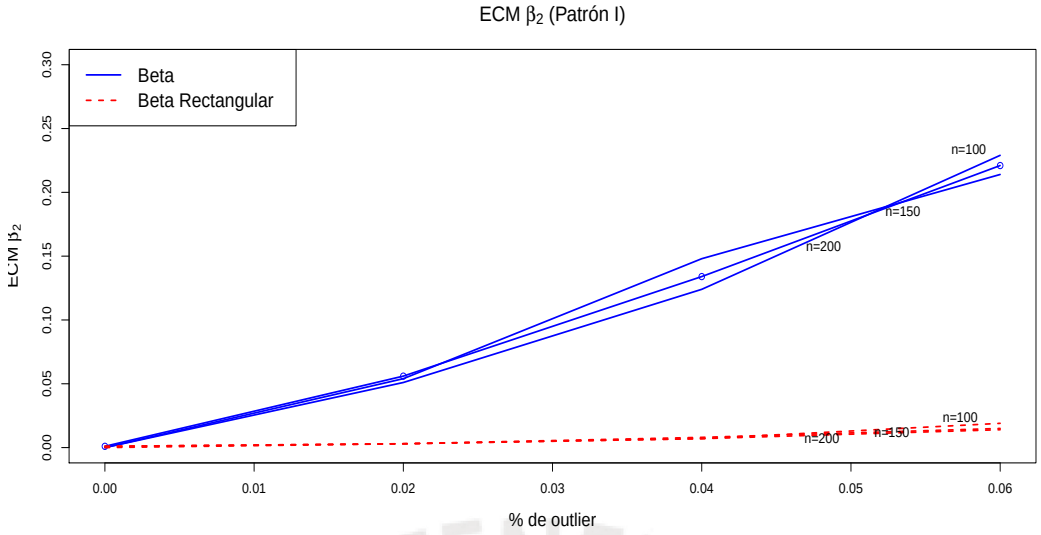


(b) ECM de β_1 siguiendo el patrón II para diferentes tamaños de n

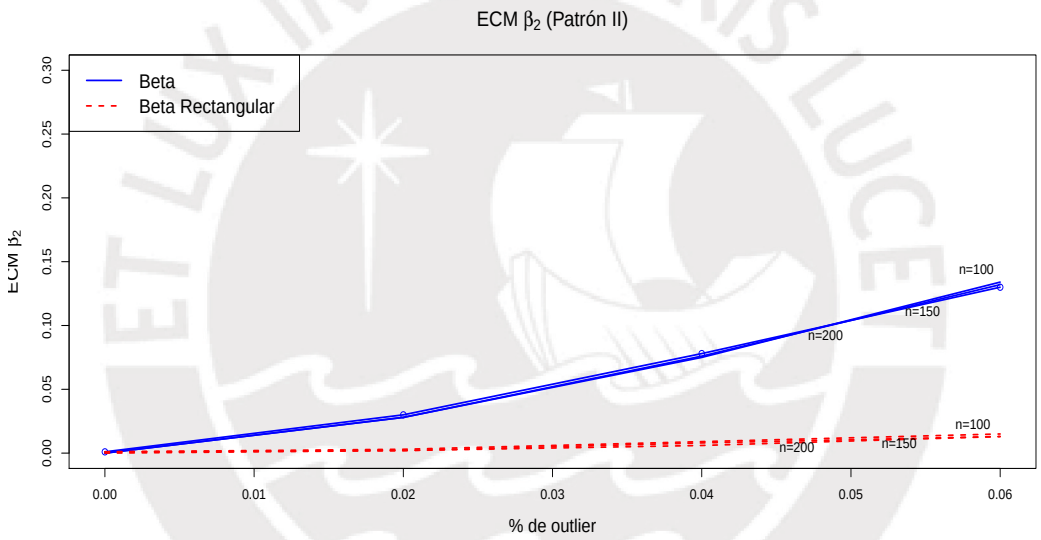


(c) ECM de β_1 siguiendo el patrón III para diferentes tamaños de n

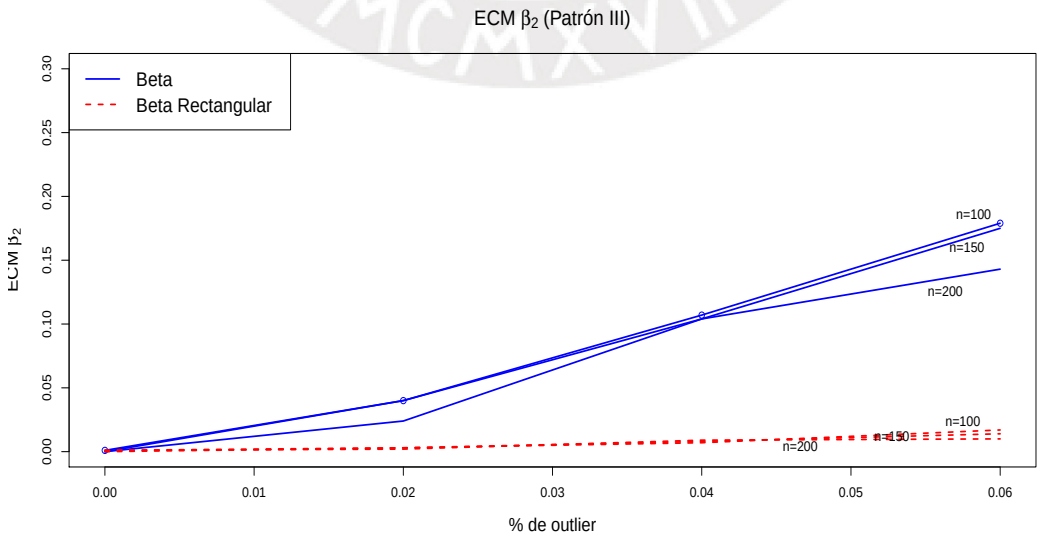
Figura 5.4: Gráficos de comparación de error cuadrático medio (ECM) para β_1 obtenidos para los modelos de regresión beta y beta rectangular considerando diferentes tamaños de n y porcentajes de valores extremos.



(a) ECM de β_2 siguiendo el patrón I para diferentes tamaños de n



(b) ECM de β_2 siguiendo el patrón II para diferentes tamaños de n



(c) ECM de β_2 siguiendo el patrón III para diferentes tamaños de n

Figura 5.5: Gráficos de comparación de error cuadrático medio (ECM) para β_2 obtenidos para los modelos de regresión beta y beta rectangular considerando diferentes tamaños de n y porcentajes de valores extremos.

Capítulo 6

Aplicación

En el presente capítulo se realizará la estimación por inferencia bayesiana de los modelos de regresión beta y beta rectangular empleando el paquete *rjags* del programa libre R, con la finalidad de comparar, para una aplicación real, ambos métodos. Los códigos se encuentran en el apéndice B.

En la aplicación referida se analizará el conjunto de datos de los 83 distritos del departamento de La Libertad según condición de pobreza y tasa de analfabetismo, 2007. Estos datos fueron extraídos del documento "La Libertad Compendio Estadístico 2012" publicado en la dirección electrónica `http://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1060/libro.pdf` y se encuentran en el apéndice C.

6.1. Estudio de caso: distritos del departamento de La Libertad según condición de pobreza y tasa de analfabetismo

6.1.1. Descripción del caso

En el Informe sobre Desarrollo Humano de las Naciones Unidas para los años 2014 y 2015 se mencionó que los países más pobres son los que tienen tasas de analfabetismo más altas. También, se menciona que los porcentajes de analfabetismo y pobreza aumentan o disminuyen conjuntamente.

Según el INEI (2015), en el año 2014, el 14 % de la población pobre de 15 o más años de edad no sabía leer ni escribir, es decir eran analfabetos. Este fenómeno afecta más a los pobres extremos ya que el 23.1 % de dicho grupo no han seguido estudios de educación primaria. Entre la población no pobre se observa una tasa de analfabetismo de 4.5 %.

A nivel de área de residencia, la tasa de analfabetismo de la población pobre del área urbana se ubicó en 8.8 % y en el área rural en 19.8 %. Entre la población no pobre la incidencia del analfabetismo en el área urbana fue de 2.9 % y en el área rural de 12.7 %.

La aplicación que se muestra en la presente sección consiste en la estimación de los parámetros de los modelos de regresión beta y beta rectangular por inferencia bayesiana. Para ello, se analizará como la tasa de analfabetismo de los distritos del departamento de La Libertad se verá influenciado por la condición de pobreza; estos datos se encuentran en el apéndice C.

6.1.2. Descripción de los datos

Se están considerando para el análisis las siguientes variables:

- Proporción total de pobres: Según el INEI (2000), la pobreza total comprende a las personas cuyos hogares tienen ingresos o consumo per cápita inferiores al costo de una canasta total de bienes y servicios mínimos esenciales. En este capítulo, se trabajará con la proporción total de pobres en los distritos del departamento de La Libertad.
- Proporción total de analfabetos: Se define como la incapacidad de leer y escribir debido a la falta de enseñanza. En este capítulo, se trabajará con la proporción total de analfabetos en los distritos del departamento de La Libertad.

En la Figura 6.1 se muestra el diagrama de dispersión para los distritos del departamento de La Libertad. Aquí se puede observar que los datos para los distritos de Condomarca y Ucuncha constituyen valores extremos para la distribución de este par de variables, porque a pesar de tener valores altos de pobreza, sus tasas de analfabetismo no son muy altas. El primero tiene un porcentaje de pobreza total de 97.5 %; sin embargo su tasa de analfabetismo es de 12.3 %. Con respecto al distrito de Ucuncha, este tiene un 91.6 % de pobreza total; pero, su tasa de analfabetismo no supera el 12.5 %.

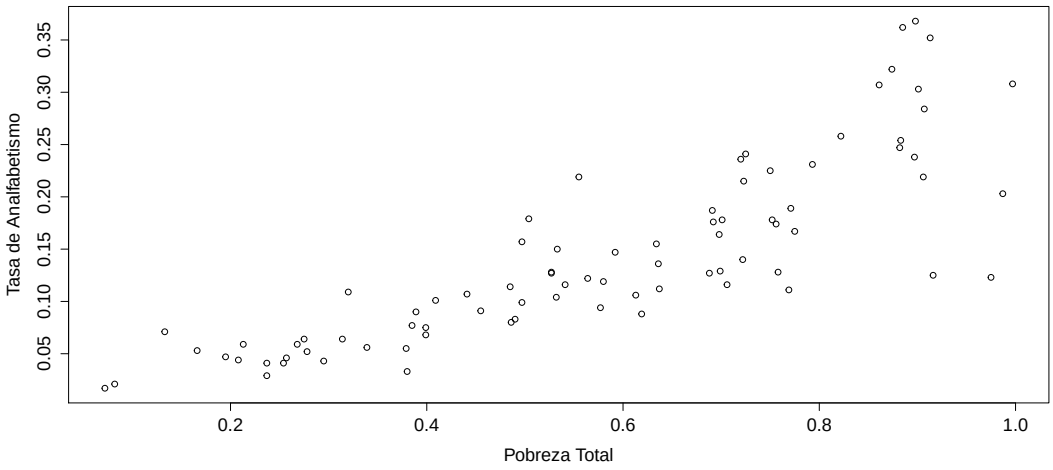


Figura 6.1: Gráfico de dispersión de tasa de analfabetismo vs pobreza total en los distritos de La Libertad.

En la Figura 6.2 se muestra el diagrama de cajas para la variable “Pobreza total” en los distritos del departamento de La Libertad. El distrito con menor cantidad de pobres es Trujillo con un 7.2 %; por otro lado, el distrito con mayor cantidad de pobres es Ongón con un 99.7 %. La media de esta variable es del 58 %, lo que indica que más de la mitad de la población de los distritos es pobre; el rango intercuantil va de 38.70 % a 75.70 %.

Por otro lado, en la Figura 6.3 se muestra el diagrama de cajas para la variable “Tasa de Analfabetismo” en los distritos del departamento de La Libertad. El distrito con menor

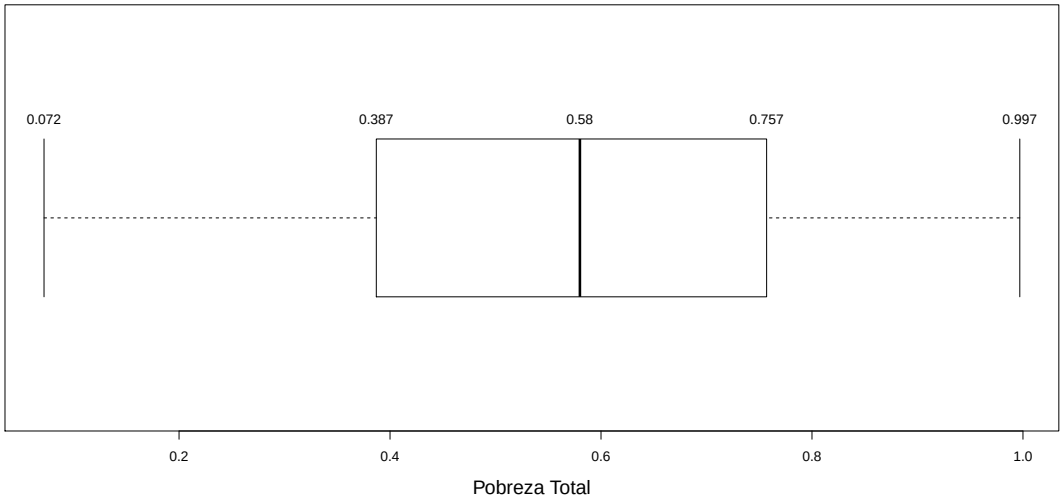


Figura 6.2: Diagrama de cajas de la proporción total de pobres en los distritos del departamento de La Libertad.

cantidad de analfabetos es Trujillo con un 1.7 %; por otro lado, el distrito con mayor cantidad de analfabetos es Sanagorán con un 36.8 %. La media de esta variable es de 14.26 %, lo que indica que cerca del 15 % de la población liberteña es analfabeta; el rango intercuantil va de 7.60 % a 18.80 %.

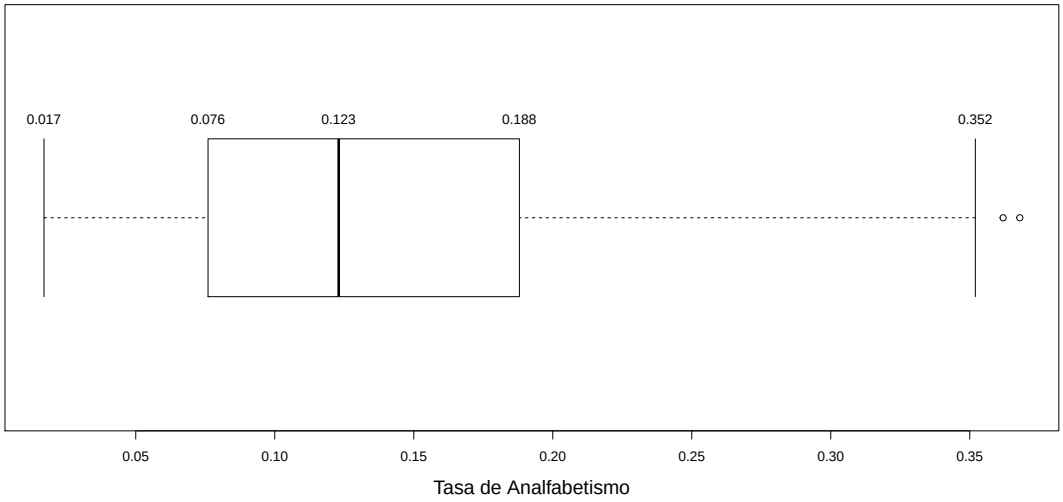


Figura 6.3: Diagrama de cajas de la tasa de analfabetismo en los distritos del departamento de La Libertad.

6.1.3. Estimación por inferencia bayesiana de los parámetros

En esta sección se muestra la estimación de los parámetros por inferencia bayesiana tanto del modelo de regresión beta como del modelo de regresión beta rectangular. La estimación

de los parámetros se realizarán en base a 4 modelos diferentes, los cuales se detallan a continuación.

1. Modelo de regresión beta con ϕ constante (modelo 1)

En primer lugar se consideró un modelo de regresión beta que tenía las siguientes características:

$$Analfabetismo_i \sim Beta(\mu_i, \phi)$$
$$logit(\mu_i) = \beta_1 + \beta_2 PobrezaTot_i$$

y distribuciones a priori no informativas dadas por $\beta_1 \sim N(0, 10^6)$, $\beta_2 \sim N(0, 10^6)$ y $\phi \sim Gamma(0.01, 0.01)$.

Se consideró que el parámetro de precisión ϕ era constante. Entonces, se realizó la estimación por MCMC de los parámetros utilizando el paquete *rjags* (ver B.1) considerando lo siguiente: 100000 iteraciones, saltos de 10 para reducir la autocorrelación y 2 cadenas; además, se descartaron las primeras 25000 observaciones. Se obtuvieron los resultados que se muestran en la Tabla 6.1. Como se puede observar en la Figura D.1, los valores simulados de los parámetros del modelo de regresión beta tienen una autocorrelación baja. Asimismo, se puede apreciar que todas las cadenas convergen; esto indica que los valores simulados son adecuados.

Parámetros	Media	Mediana	Desv. Est.	Intervalo de Credibilidad 2.5 %	97.5 %
β_1	-3.38	-3.38	0.13	-3.63	-3.13
β_2	2.55	2.55	0.18	2.19	2.89
ϕ	63.27	62.68	9.92	45.33	84.02

Tabla 6.1: Estimación de los parámetros del modelo 1 considerando μ variable y ϕ constante. Se puede apreciar que β_1 y β_2 son significativos ya que en sus intervalos de credibilidad no contienen al cero.

El valor estimado de la constante β_1 es -3.38 y el valor estimado de β_2 es 2.55, lo que indica que a medida que aumenta el nivel de Pobreza, aumenta el nivel de Analfabetismo. Cuando la tasa de Pobreza Total es igual a 0, se espera que la tasa de analfabetismo estimada sea de 3.29 %.

2. Modelo de regresión beta con ϕ variable (modelo 2)

Se consideró un modelo de regresión beta que tenía las siguientes características:

$$Analfabetismo_i \sim Beta(\mu_i, \phi_i)$$
$$logit(\mu_i) = \beta_1 + \beta_2 PobrezaTot_i$$
$$log(\phi_i) = -\delta_1 - \delta_2 PobrezaTot_i$$

y distribuciones a priori no informativas dadas por $\beta_1 \sim N(0, 10^6)$, $\beta_2 \sim N(0, 10^6)$, $\delta_1 \sim N(0, 10^6)$ y $\delta_2 \sim N(0, 10^6)$.

Parámetros	Media	Mediana	Desv. Est.	Intervalo de Credibilidad	
				2.5 %	97.5 %
β_1	-3.58	-3.58	0.10	-3.77	-3.37
β_2	2.83	2.84	0.17	2.50	3.17
δ_1	-5.77	-5.78	0.38	-6.47	-4.99
δ_2	2.45	2.45	0.59	1.28	3.60

Tabla 6.2: Estimación de los parámetros del modelo 2 considerando μ y ϕ variables. Se puede apreciar que β_1, β_2, δ_1 y δ_2 son significativos ya que en sus intervalos de credibilidad no contienen al cero.

Entonces, se realizó la estimación por MCMC de los parámetros utilizando el paquete *rjags* (ver B.1) considerando lo siguiente: 100000 iteraciones, saltos de 10 para reducir la autocorrelación y 2 cadenas; además, se descartaron las primeras 25000 observaciones. Se obtuvieron los resultados que se muestran en la Tabla 6.2. Como se puede observar en la Figura D.2, los valores simulados de los parámetros del modelo de regresión beta tienen una concordancia muy buena. También, se puede apreciar que todas las cadenas convergen.

Se observa que el valor estimado de la constante β_1 es -3.58 y el valor estimado de β_2 es 2.83, lo que indica que a medida que aumenta el nivel de Pobreza, aumenta el nivel de Analfabetismo. Por otro lado, el valor de δ_1 es la constante negativa, que al entrar al modelo tiene un efecto positivo; asimismo, el valor de δ_2 es positivo y al entrar al modelo tiene un efecto negativo sobre la Pobreza Total; por ello, el valor de ϕ_i disminuye, haciendo que la varianza del Analfabetismo aumente. Cuando la tasa de Pobreza Total es igual a 0, la tasa de analfabetismo es de 2.71 %.

3. Modelo de regresión beta rectangular con ϕ constante (modelo 3)

Se consideró un modelo de regresión beta rectangular que tenía las siguientes características:

$$Analfabetismo_i \sim BR(\gamma_i, \phi, \alpha)$$
$$logit(\gamma_i) = \beta_1 + \beta_2 PobrezaTot_i$$

y distribuciones a priori no informativas dadas por $\beta_1 \sim N(0, 10^6)$, $\beta_2 \sim N(0, 10^6)$, $\phi \sim Gamma(0.01, 0.01)$ y $\alpha \sim U(0, 1)$.

Se consideró que la media γ dependía de covariables; mientras que ϕ y α eran constantes. Luego, se realizó la estimación por MCMC de los parámetros utilizando el paquete *rjags* (ver B.2) considerando lo siguiente: 100000 iteraciones, saltos de 10 para reducir la autocorrelación y 2 cadenas; además, se descartaron las primeras 25000 observaciones. Se obtuvieron los resultados que se muestran en la Tabla 6.3. Como se puede observar en la Figura D.3, los valores simulados de los parámetros del modelo de regresión beta rectangular tienen una concordancia muy buena; esto significa que el proceso de simulación ha sido adecuado. Asimismo, se puede apreciar que todas las cadenas convergen.

El valor estimado de la constante β_1 es -3.39 y el valor estimado de β_2 es 2.75, lo que indica que a medida que aumenta el nivel de Pobreza, aumenta el nivel de Analfabetismo. Cuando la tasa de Pobreza Total es igual a 0, la tasa de analfabetismo es de 3.26 %.

Parámetros	Media	Mediana	Desv. Est.	Intervalo de Credibilidad 2.5 %	97.5 %
β_1	-3.39	-3.39	0.15	-3.67	-3.09
β_2	2.75	2.76	0.19	2.36	3.12
α	0.13	0.12	0.07	0.01	0.30
ϕ	86.43	85.31	17.58	55.55	124.58

Tabla 6.3: Estimación de los parámetros del modelo 3 considerando γ variable, ϕ y α constantes. Se puede apreciar que β_1 y β_2 son significativos ya que en sus intervalos de credibilidad no contienen al cero.

4. Modelo de regresión beta rectangular con ϕ variable (modelo 4)

Se consideró un modelo de regresión beta rectangular que tenía las siguientes características:

$$Analfabetismo_i \sim BR(\gamma_i, \phi_i, \alpha)$$
$$logit(\gamma_i) = \beta_1 + \beta_2 PobrezaTot_i$$
$$log(\phi_i) = -\delta_1 - \delta_2 PobrezaTot_i$$

y distribuciones a priori no informativas dadas por $\beta_1 \sim N(0, 10^6)$, $\beta_2 \sim N(0, 10^6)$, $\delta_1 \sim N(0, 10^6)$, $\delta_2 \sim N_N(0, 10^6)$ y $\alpha \sim U(0, 1)$.

Se consideró que γ y ϕ dependían de covariables; mientras que α se mantenía constante. Luego, se realizó la estimación por MCMC de los parámetros utilizando el paquete *rjags* (ver B.2) considerando lo siguiente: 100000 iteraciones, saltos de 10 para reducir la autocorrelación y 2 cadenas; además, se descartaron las primeras 25000 observaciones. Se obtuvieron los resultados que se muestran en la Tabla 6.4. Como se puede observar en la Figura D.4 y la Figura D.5 , los valores simulados de los parámetros del modelo de regresión beta rectangular tienen una concordancia muy buena. Nuevamente, significa que el proceso de simulación ha sido adecuado. También, se puede apreciar que todas las cadenas convergen.

Parámetros	Media	Mediana	Desv. Est.	Intervalo de Credibilidad 2.5 %	97.5 %
β_1	-3.52	-3.52	0.11	-3.73	-3.28
β_2	2.82	2.82	0.17	2.49	3.15
δ_1	-5.73	-5.74	0.39	-6.44	-4.93
δ_2	2.33	2.35	0.63	1.02	3.53
α	0.06	0.04	0.05	0.00	0.19

Tabla 6.4: Estimación de los parámetros del modelo 4 considerando γ variable, ϕ variable y α constante. Se puede apreciar que β_1, β_2, δ_1 y δ_2 son significativos ya que en sus intervalos de credibilidad no contienen al cero.

Se observa que el valor estimado de la constante β_1 es -3.52 y el valor estimado de β_2 es 2.82, lo que indica que a medida que aumenta el nivel de Pobreza, aumenta el nivel de Analfabetismo. Por otro lado, el valor de δ_1 es la constante negativa, que al entrar al modelo tiene un efecto positivo; asimismo, el valor de δ_2 es positivo y al entrar al modelo tiene un efecto negativo sobre la Pobreza Total; por ello, el valor de ϕ_i disminuye, haciendo que la varianza del Analfabetismo aumente. Cuando la tasa de Pobreza Total es igual a 0, la tasa

de analfabetismo es de 2.87 %.

6.1.4. Resultados de la aplicación

En la presente sección se comparan los criterios de información $EAIC$, $EBIC$ y DIC para cada modelo de regresión beta y beta rectangular. En la Tabla 6.5 se pueden apreciar dichos valores calculados.

Criterios de información	Modelos de regresión			
	Beta		Beta rectangular	
	Modelo 1	Modelo 2	Modelo 3	Modelo 4
$EAIC$	-296.43	-280.22	-296.66	-280.71
$EBIC$	-296.67	-280.54	-296.98	-281.12
DIC	-299.40	-284.42	-300.86	-286.91

Tabla 6.5: Criterios de información para los diferentes modelos de regresión beta y beta rectangular.

Los criterios de información indican que entre los modelos de regresión beta, el mejor es el que considera a ϕ constante (Modelo 1); por otro lado, entre los modelos de regresión beta rectangular, el mejor es también el que considera a ϕ constante (Modelo 3). Este último es el que tiene un mejor ajuste de entre todos los modelos utilizados según el análisis de criterios.

6.1.5. Comparación de los mejores modelos de regresión beta y beta rectangular luego de retirar observaciones atípicas

En esta sección se muestra una comparación, luego de retirar las observaciones atípicas, entre el Modelo 3 que fue el que obtuvo mejor ajuste de todos los modelos con el Modelo 1 porque para ambos el parámetro ϕ es constante . En el panel izquierdo de la Figura 6.4 se muestra el gráfico de dispersion donde se marcaron los distritos de Condomarca y Ucuncha porque constituían valores extremos para la distribución de los datos; además, se visualizan los modelos de regresión beta y beta rectangular estimados.

Primero, se realizó la estimación por MCMC de los parámetros utilizando un modelo de regresión beta con ϕ constante; para ello, considerando lo siguiente: 100000 iteraciones, saltos de 10 para reducir la autocorrelación y 2 cadenas; además, se descartaron las primeras 25000 observaciones. Se obtienen los resultados que se muestran en la Tabla 6.6.

Parámetros	Media	Mediana	Desv. Est.	Intervalo de Credibilidad	
				2.5 %	97.5 %
β_1	-3.54	-3.54	0.11	-3.77	-3.32
β_2	2.83	2.83	0.16	2.51	3.15
ϕ	85.97	85.19	13.77	60.77	115.02

Tabla 6.6: Estimación de los parámetros considerando un modelo de regresión beta con μ variable y ϕ constante. Se puede apreciar que β_1, β_2 y ϕ son significativos ya que en sus intervalos de confianza no se considera al cero.

Luego, se realizó la estimación por MCMC de los parámetros utilizando un modelo de regresión beta rectangular con ϕ constante; considerando lo siguiente: 100000 iteraciones, saltos de 10 para reducir la autocorrelación y 2 cadenas; además, se descartaron las primeras 25000 observaciones. Se obtienen los resultados que se muestran en la Tabla 6.7.

Parámetros	Media	Mediana	Desv. Est.	Intervalo de Credibilidad	
				2.5 %	97.5 %
β_1	-3.49	-3.49	0.13	-3.74	-3.23
β_2	2.82	2.82	0.17	2.49	3.15
α	0.05	0.04	0.05	0.00	0.19
ϕ	88.05	86.99	14.61	62.38	119.49

Tabla 6.7: Estimación de los parámetros considerando un modelo de regresión beta rectangular con γ variable y ϕ constante. Se puede apreciar que β_1, β_2, α y ϕ son significativos ya que en sus intervalos de confianza no se considera al cero.

Finalmente, se comparan los criterios de información $EAIC$, $EBIC$ y DIC para los modelos evaluados.

Criterios de información	Modelos de regresión	
	Beta	Beta rectangular
	Modelo 1	Modelo 3
$EAIC$	-313.83	-310.13
$EBIC$	-314.11	-310.50
DIC	-316.70	-315.07

Tabla 6.8: Criterios de información para el modelo de regresión beta y beta rectangular utilizando datos sin valores atípicos

Los criterios de información indican que el mejor modelo, cuando se retiran los valores atípicos, es el de regresión beta que emplea ϕ constante (modelo 1). También, en el panel izquierdo de la Figura 6.4 se muestran los datos con dos observaciones atípicas y los modelos de regresión beta y beta rectangular, luego en el panel derecho se aprecian ambos modelos reajustados después de retirar estos valores atípicos. Esto indica que no existe mucha diferencia en la estimación de la media de la variable Analfabetismo cuando no hay valores atípicos; sin embargo, dichos valores pueden tener un impacto en las estimaciones del modelo de regresión beta, mientras que este comportamiento no se observó en el modelo de regresión beta rectangular.

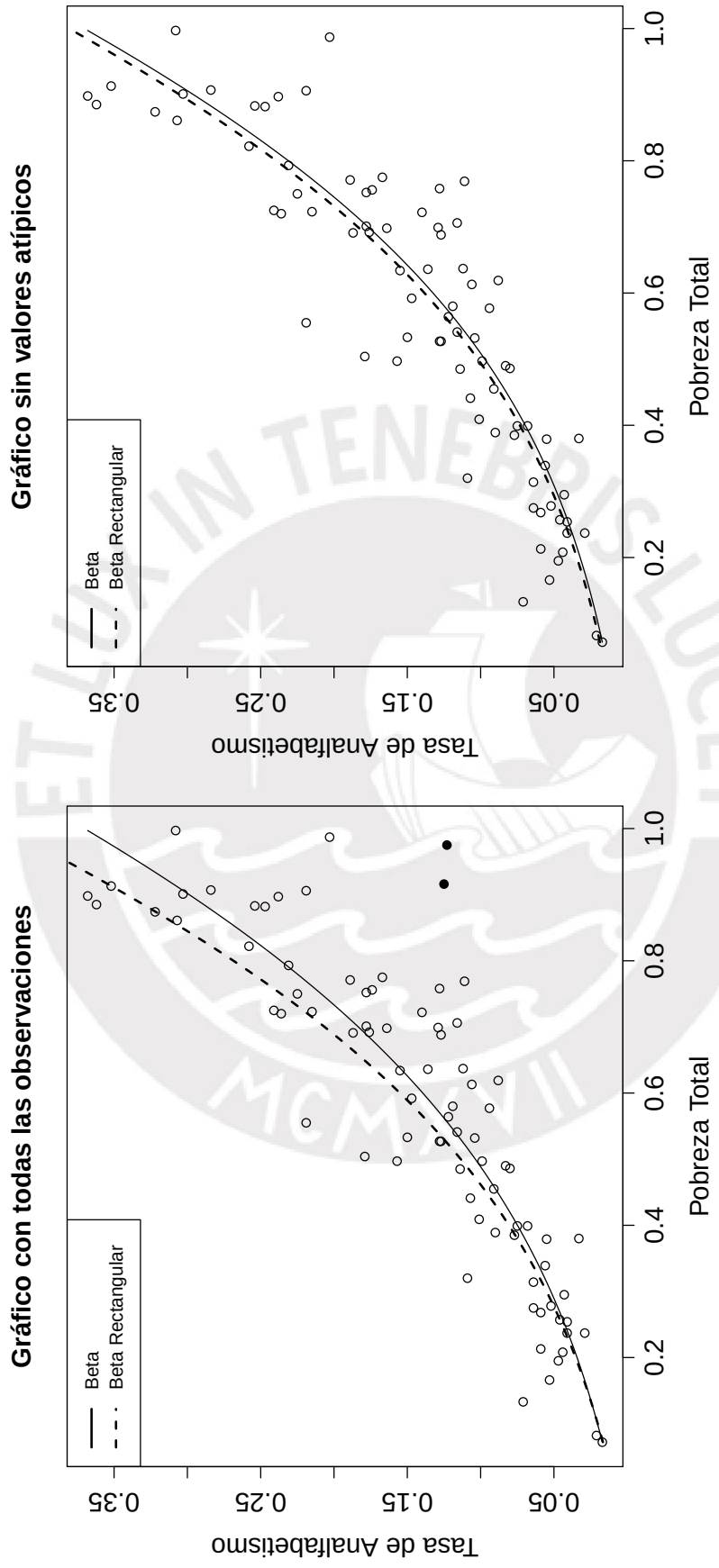


Figura 6.4: Gráfico de dispersión y modelo ajustado de la tasa de analfabetismo vs pobreza total en los distritos de La Libertad considerando todas las observaciones (panel izquierdo) y sin valores atípicos (panel derecho).

Capítulo 7

Conclusiones y sugerencias

7.1. Conclusiones

El estudio de simulación realizado demostró que el modelo de regresión beta rectangular resulta ser más apropiado que el modelo de regresión beta ante la presencia de observaciones atípicas o extremas. En este estudio se vió como este modelo mostraba mejores indicadores de ajuste (menores valores de sesgo y ECM); así como un menor valor de DIC para los escenarios estudiados.

Mediante la aplicación realizada al conjunto de datos de distritos del departamento de La Libertad según condición de pobreza y tasa de analfabetismo, se verificó que el modelo de regresión beta rectangular presentó un mejor ajuste cuando los datos presentan valores atípicos o extremos. Los datos muestran que los distritos de Condomarca y Ucuncha constituyen valores extremos para la distribución de los datos, porque a pesar de tener valores altos de pobreza, sus tasas de analfabetismo no son muy altas. El mejor ajuste fue obtenido mediante el modelo de regresión beta rectangular, que busca explicar la tasa de analfabetismo con la covariable de porcentaje de pobreza en donde el parámetro de precisión ϕ se considera constante (modelo 3). Asimismo, cuando se retiran los valores extremos del conjunto de datos, no existe mucha diferencia entre este y el modelo beta; sin embargo, dichos valores pueden tener un impacto en las estimaciones del modelo de regresión beta, por lo que el modelo de regresión beta rectangular, se podría catalogar como más seguro de utilizar y replicar para otras aplicaciones en donde se tenga valores extremos o atípicos.

7.2. Sugerencias para investigaciones futuras

Algunos temas que pueden ser estudiados a profundidad son los siguientes:

- Realizar estudios de sensibilidad considerando otras distribuciones a priori para los parámetros y estudios de simulación con más variables predictoras.
- Desarrollar técnicas de detección de puntos influyentes para el modelo de regresión beta rectangular.
- Extender el modelo beta rectangular para el análisis de datos longitudinales.

Apéndice A

Código en *rjags* para realizar simulaciones de datos considerando los modelos de regresión beta y beta rectangular

A continuación, se muestran los códigos en R que se utilizaron en la simulación de los modelos de regresión beta y beta rectangular para comparar el sesgo y error cuadrático medio de los parámetros obtenidos. Para ello, se utilizaron los *software* libres R en su versión 3.2.2 y *JAGS* en su versión 3.4.0. Asimismo, para los ajustes de los modelos bajo la perspectiva bayesiana se usó la función *jags.model* del paquete *rjags* de R.

A.1. Código *R* para la simulación del modelo de regresión beta

A.1.1. Simulación del modelo de regresión beta sin data contaminada

```
library(boot)
library(coda)
library(rjags)

M=100
beta1.B= numeric(M)
beta2.B= numeric(M)
phi.B= numeric(M)

##Simulación del conjunto de datos sin contaminar##
n=100

x=runif(n,-4,4)
beta1or=0.6
beta2or=1
phior=40
mu1=inv.logit(beta1or+beta2or*x)

##Modelo de regresión beta##
```

```
model_string1= "model {
for (i in 1:n) {
y[i] ~ dbeta(a[i],b[i])
a[i]<-mu[i]*phi
b[i]<-phi*(1-mu[i])
logit(mu[i])<-beta1 + beta2*x[i]
}
beta1 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
phi ~ dgamma(0.01,0.01)
}"

for(h in 1:M){
y=rbeta(n,mu1*phior,(1-mu1)*phior)

##Aplicación de rjags##
jags <- jags.model(textConnection(model_string1),
                   data = list('x' = x,
                               'y' = y,
                               'n' = n),
                   n.chains = 2,
                   n.adapt = 5000)

update(jags, 25000)

samp.jags=jags.samples(jags, c('beta1', 'beta2','phi'),100000,thin=10)

samps.coda=coda.samples(jags,
                        c('beta1','beta2','phi'),
                        n.iter=20000,
                        thin=5
                        )
head(samps.coda)
summary(samps.coda)
#plot(samps.coda[[1]][,c("beta1","beta2","phi")])

valbeta1=as.matrix(samps.coda[,1])
valbeta2=as.matrix(samps.coda[,2])
valphi=as.matrix(samps.coda[,3])

beta1.B[h]=mean(valbeta1)
beta2.B[h]=mean(valbeta2)
```

```
phi.B[h]=mean(valphi)
}

##Cálculo del sesgo##
sesgobeta1=mean(beta1.B)-0.6
sesgobeta2=mean(beta2.B)-1
sesgophi=mean(phi.B)-40
sesgobeta1
sesgobeta2
sesgophi

##Cálculo del ECM##
ECMbeta1=mean((beta1.B-0.6)**2)
ECMbeta2=mean((beta2.B-1)**2)
ECMphi=mean((phi.B-40)**2)
ECMbeta1
ECMbeta2
ECMphi

##Cálculo del Error Estándar##
EEbeta1=sd(beta1.B)
EEbeta2=sd(beta2.B)
EEphi=sd(phi.B)
EEbeta1
EEbeta2
EEphi

##Valor DIC##
dic.pD <- dic.samples(jags, 100000, "pD")
dic.pD
```

A.1.2. Simulación del modelo de regresión beta con data contaminada

```
library(boot)
library(coda)
library(rjags)

M=100
beta1.B=numeric(M)
beta2.B=numeric(M)
phi.B=numeric(M)

##Simulación del conjunto de datos##
```

```
n=100

x=runif(n,-4,4)
beta1or=0.6
beta2or=1
phior=40
mu1=inv.logit(beta1or+beta2or*x)

##Modelo de regresión beta##
model_string1= "model {
for (i in 1:n) {
yprim[i] ~ dbeta(a[i],b[i])
a[i]<-mu[i]*phi
b[i]<-phi*(1-mu[i])
logit(mu[i])<-beta1 + beta2*x[i]
}
beta1 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
phi ~ dgamma(0.01,0.01)
}"

for(h in 1:M){
y=rbeta(n,mu1*phior,(1-mu1)*phior)

##Contaminando la data##
r=0.02
cant=r*n
s1=sample(y[y<0.05],cant)
posicion=match(s1,y)
ycont=s1+0.9
yprim=replace(y,posicion,ycont)
#plot(x,yprim)

##Aplicación de rjags##
jags <- jags.model(textConnection(model_string1),
                   data = list('x' = x,
                               'yprim' = yprim,
                               'n' = n),
                   n.chains = 2,
                   n.adapt = 5000)

update(jags, 25000)
```

```
samp.jags=jags.samples(jags, c('beta1', 'beta2','phi'),100000,thin=10)

samps.coda=coda.samples(jags,
                        c('beta1','beta2','phi'),
                        n.iter=20000,
                        thin=5
                        )
head(samps.coda)
summary(samps.coda)
#plot(samps.coda[[1]][,c("beta1","beta2","phi")])

valbeta1=as.matrix(samps.coda[,1])
valbeta2=as.matrix(samps.coda[,2])
valphi=as.matrix(samps.coda[,3])

beta1.B[h]=mean(valbeta1)
beta2.B[h]=mean(valbeta2)
phi.B[h]=mean(valphi)
}

##Cálculo del sesgo##
sesgobeta1=mean(beta1.B)-0.6
sesgobeta2=mean(beta2.B)-1
sesgophi=mean(phi.B)-40
sesgobeta1
sesgobeta2
sesgophi

##Cálculo del ECM##
ECMbeta1=mean((beta1.B-0.6)**2)
ECMbeta2=mean((beta2.B-1)**2)
ECMphi=mean((phi.B-40)**2)
ECMbeta1
ECMbeta2
ECMphi

##Cálculo del Error Estándar##
EEbeta1=sd(beta1.B)
EEbeta2=sd(beta2.B)
EEphi=sd(phi.B)
EEbeta1
```

```
EEbeta2
EEphi

##Valor DIC##
dic.pD <- dic.samples(jags, 100000, "pD")
dic.pD
```

A.2. Código R para la simulación del modelo de regresión beta rectangular

A.2.1. Simulación del modelo de regresión beta rectangular sin data contaminada

```
library(boot)
library(coda)
library(rjags)

M=100
beta1.BR= numeric(M)
beta2.BR= numeric(M)
alpha.BR= numeric(M)
phi.BR= numeric(M)

##Simulación del conjunto de datos sin contaminar##
n=100

x2=runif(n,-4,4)
beta1or=0.6
beta2or=1
alphaor=0.4
phior=40
gamma1=inv.logit(beta1or+beta2or*x2)

##Modelo de regresión beta rectangular##
model_string2= "model {
       for (i in 1:n) {
y2[i] ~ dbeta(a[i,u[i]],b[i,u[i]])
v[i] ~ dbern(theta[i])
u[i] <- v[i]+1
a[i,1] <- mu[i]*phi
b[i,1] <-(1-mu[i])*phi
a[i,2] <- 1
b[i,2] <- 1
```

```
logit(gamma[i]) <- beta1.star+beta2*(x2[i]-mean(x2[]))
theta[i]<-alpha*(1-2*abs(gamma[i]-0.5))
mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
}

##Definiendo la distribución de los parametros##
beta2 ~dnorm (0,0.0001)
beta1.star ~dnorm (0,0.0001)
beta1<-beta1.star-beta2*mean(x2[])
alpha ~ dunif(0,1)
phi~ dgamma(0.01,0.01)
}"

for(h in 1:M){
y2=rbeta(n,gamma1*phior,(1-gamma1)*phior)

##Aplicación de rjags##
jags2 <- jags.model(textConnection(model_string2),
                    data = list('x2' = x2,
                                'y2' = y2,
                                'n' = n),
                    n.chains = 2,
                    n.adapt = 5000)

update(jags2, 25000)

sam.jags2=jags.samples(jags2, c('alpha','beta1','beta2','phi'),100000,thin=10)

samps.coda2=coda.samples(jags2,
                         c('alpha','beta1','beta2','phi'),
                         n.iter=20000,
                         thin=5
                         )

head(samps.coda2)
summary(samps.coda2)
#plot(samps.coda2[[1]][,c("alpha","beta1","beta2","phi")])

valalpha=as.matrix(samps.coda2[,1])
valbeta1r=as.matrix(samps.coda2[,2])
valbeta2r=as.matrix(samps.coda2[,3])
valphi=as.matrix(samps.coda2[,4])
```

```
alpha.BR[h]=mean(valalpha)
beta1.BR[h]=mean(valbeta1r)
beta2.BR[h]=mean(valbeta2r)
phi.BR[h]=mean(valphi)
}

##Cálculo del sesgo##
sesgobeta1r=mean(beta1.BR)-0.6
sesgobeta2r=mean(beta2.BR)-1
sesgoalpha=mean(alpha.BR)-0.4
sesgophi=mean(phi.BR)-40
sesgobeta1r
sesgobeta2r
sesgoalpha
sesgophi

##Cálculo del ECM##
ECMbeta1r=mean((beta1.BR-0.6)**2)
ECMbeta2r=mean((beta2.BR-1)**2)
ECMalpha=mean((alpha.BR-1)**2)
ECMphi=mean((phi.BR-40)**2)
ECMbeta1r
ECMbeta2r
ECMalpha
ECMphi

##Cálculo del Error Estándar##
EEbeta1r=sd(beta1.BR)
EEbeta2r=sd(beta2.BR)
EEalpha=sd(alpha.BR)
EEphi=sd(phi.BR)
EEbeta1r
EEbeta2r
EEalpha
EEphi

##Valor DIC##
dic.pD <- dic.samples(jags, 100000, "pD")
dic.pD
```

A.2.2. Simulación del modelo de regresión beta rectangular con data contaminada

```
library(boot)
library(coda)
library(rjags)

M=100
beta1.BR= numeric(M)
beta2.BR= numeric(M)
alpha.BR= numeric(M)
phi.BR= numeric(M)

##Simulación del conjunto de datos sin contaminar##
n=100

x2=runif(n,-4,4)
beta1or=0.6
beta2or=1
alphaor=0.4
phior=40
gamma1=inv.logit(beta1or+beta2or*x2)

##Modelo de regresión beta rectangular##
model_string2= "model {
       for (i in 1:n) {
yprim2[i] ~ dbeta(a[i,u[i]],b[i,u[i]])
v[i] ~ dbern(theta[i])
u[i] <- v[i]+1
a[i,1] <- mu[i]*phi
b[i,1] <-(1-mu[i])*phi
a[i,2] <- 1
b[i,2] <- 1

logit(gamma[i]) <- beta1.star+beta2*(x2[i]-mean(x2[]))
theta[i]<-alpha*(1-2*abs(gamma[i]-0.5))
mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
}

##Definiendo la distribución de los parámetros##
beta2 ~dnorm (0,0.0001)
beta1.star ~dnorm (0,0.0001)
beta1<-beta1.star-beta2*mean(x2[])
```

```
alpha ~ dunif(0,1)
phi~ dgamma(0.01,0.01)
}"

for(h in 1:M){
y2=rbeta(n,gamma1*phior,(1-gamma1)*phior)

##Contaminando la data##
r2=0.02
cant2=r2*n
s3=sample(y2[y2<0.05],cant2)
posicion3=match(s3,y2)
ycont2=s3+0.90
yprim2=replace(y2,posicion3,ycont2)
#plot(x2,yprim2)

##Aplicación de rjags##
jags2 <- jags.model(textConnection(model_string2),
                    data = list('x2' = x2,
                                'yprim2' = yprim2,
                                'n' = n),
                    n.chains = 2,
                    n.adapt = 5000)

update(jags2, 25000)

sam.jags2=jags.samples(jags2, c('alpha','beta1','beta2','phi'),100000,thin=10)

samps.coda2=coda.samples(jags2,
                         c('alpha','beta1','beta2','phi'),
                         n.iter=20000,
                         thin=5
                         )

head(samps.coda2)
summary(samps.coda2)
#plot(samps.coda2[[1]][,c("alpha","beta1","beta2","phi")])

valalpha=as.matrix(samps.coda2[,1])
valbeta1r=as.matrix(samps.coda2[,2])
valbeta2r=as.matrix(samps.coda2[,3])
valphi=as.matrix(samps.coda2[,4])
```

```
alpha.BR[h]=mean(valalpha)
beta1.BR[h]=mean(valbeta1r)
beta2.BR[h]=mean(valbeta2r)
phi.BR[h]=mean(valphi)
}

##Cálculo del sesgo##
sesgobeta1r=mean(beta1.BR)-0.6
sesgobeta2r=mean(beta2.BR)-1
sesgoalpha=mean(alpha.BR)-0.4
sesgophi=mean(phi.BR)-40
sesgobeta1r
sesgobeta2r
sesgoalpha
sesgophi

##Cálculo del ECM##
ECMbeta1r=mean((beta1.BR-0.6)**2)
ECMbeta2r=mean((beta2.BR-1)**2)
ECMalpha=mean((alpha.BR-1)**2)
ECMphi=mean((phi.BR-40)**2)
ECMbeta1r
ECMbeta2r
ECMalpha
ECMphi

##Cálculo del Error Estándar##
EEbeta1r=sd(beta1.BR)
EEbeta2r=sd(beta2.BR)
EEalpha=sd(alpha.BR)
EEphi=sd(phi.BR)
EEbeta1r
EEbeta2r
EEalpha
EEphi

##Valor DIC##
dic.pD <- dic.samples(jags, 100000, "pD")
dic.pD
```

Apéndice B

Programas utilizados en las estimaciones de modelos de regresión beta y beta rectangular

B.1. Código *rjags* de estimación del modelo de regresión beta con la data de distritos del departamento de La Libertad

1. Modelo de regresión beta cuando μ es variable y ϕ es constante

```
model {
for (i in 1:n) {
y[i] ~ dbeta(a[i],b[i])
a[i]<-mu[i]*phi
b[i]<-phi*(1-mu[i])
logit(mu[i])<-beta1 + beta2*x[i]
}

#Definiendo la distribución de los parámetros#
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
phi ~ dgamma(0.01,0.01)
}
```

2. Modelo de regresión beta cuando μ y ϕ son variables

```
model{
for (i in 1:n) {
y[i] ~ dbeta(a[i],b[i])
a[i]<-mu[i]*phi[i]
b[i]<-phi[i]*(1-mu[i])
logit(mu[i])<- beta1 + beta2*x[i]
log(phi[i]) <- -delta1 - delta2*x[i]
}
```

```
#Definiendo la distribución de los parámetros#
beta1 ~ dnorm(0.0,1.0E-6)
beta2 ~ dnorm(0.0,1.0E-6)
delta1 ~ dnorm(0.0,1.0E-6)
delta2 ~ dnorm(0.0,1.0E-6)
}
```

B.2. Código *rjags* de estimación del modelo de regresión beta rectangular con la data de distritos del departamento de La Libertad

1. Modelo de regresión beta rectangular cuando γ es variable y ϕ es constante

```
model {
       for (i in 1:n) {
y[i] ~ dbeta(a[i,u[i]],b[i,u[i]])
v[i] ~ dbern(theta[i])
u[i] <- v[i]+1
a[i,1] <- mu[i]*phi
b[i,1] <-(1-mu[i])*phi
a[i,2] <- 1
b[i,2] <- 1

logit(gamma[i]) <- beta1+beta2*x[i]
theta[i]<-alpha*(1-2*abs(gamma[i]-0.5))
mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
}

#Definiendo la distribución de los parámetros
beta1 ~dnorm (0.0,1.0E-6)
beta2 ~dnorm (0.0,1.0E-6)
alpha ~ dunif(0,1)
phi~ dgamma(0.01,0.01)
}
```

2. Modelo de regresión beta rectangular cuando γ y ϕ son variables

```
model {
       for (i in 1:n) {
y[i] ~ dbeta(a[i,u[i]],b[i,u[i]])
v[i] ~ dbern(theta[i])
u[i] <- v[i]+1
a[i,1] <- mu[i]*phi[i]
```

```
b[i,1] <-(1-mu[i])*phi[i]
a[i,2] <- 1
b[i,2] <- 1

logit(gamma[i]) <- beta1+beta2*x[i]
log(phi[i]) <- -delta1-delta2*x[i]
theta[i]<-alpha*(1-2*abs(gamma[i]-0.5))
mu[i]<-(gamma[i]-0.5*theta[i])/(1-theta[i])
}

#Definiendo la distribución de los parámetros
beta1 ~dnorm (0.0,1.0E-6)
beta2 ~dnorm (0.0,1.0E-6)
delta1 ~dnorm (0.0,1.0E-6)
delta2 ~dnorm (0.0,1.0E-6)
alpha ~ dunif(0,1)
}
```

Apéndice C

Datos empleados en la aplicación

Distritos del departamento de La Libertad	Población	Pobres (%)			No Pobres (%)	Tasa de Analfabetismo (%)
		Total	Extremos	No extremos		
Ongón	1,784	99.7 %	97.2 %	2.5 %	0.3 %	30.8 %
Bambamarca	3,826	98.7 %	92.4 %	6.3 %	1.3 %	20.3 %
Condomarca	2,403	97.5 %	83.3 %	14.2 %	2.5 %	12.3 %
Ucuncha	1,023	91.6 %	58.6 %	33.0 %	8.4 %	12.5 %
Sartimbamba	13,328	91.3 %	58.2 %	33.1 %	8.7 %	35.2 %
Curgos	8,621	90.7 %	54.0 %	36.7 %	9.3 %	28.4 %
Sinsicap	8,729	90.6 %	55.4 %	35.2 %	9.4 %	21.9 %
Chugay	18,163	90.1 %	55.2 %	34.9 %	9.9 %	30.3 %
Sanagorán	13,682	89.8 %	53.6 %	36.2 %	10.2 %	36.8 %
Sitabamba	3,907	89.7 %	51.5 %	38.2 %	10.3 %	23.8 %
Marcabal	15,604	88.5 %	47.5 %	41.0 %	11.5 %	36.2 %
Huayo	4,366	88.3 %	50.5 %	37.8 %	11.7 %	25.4 %
Cochorco	9,222	88.2 %	48.2 %	40.0 %	11.8 %	24.7 %
Chillia	12,686	87.4 %	48.8 %	38.6 %	12.6 %	32.2 %
Sarín	9,649	86.1 %	46.7 %	39.4 %	13.9 %	30.7 %
Huancaspata	6,691	82.2 %	38.7 %	43.5 %	17.8 %	25.8 %
Tayabamba	14,521	79.3 %	37.6 %	41.7 %	20.7 %	23.1 %
Carabamba	7,292	77.5 %	32.1 %	45.4 %	22.5 %	16.7 %
Usquil	27,722	77.1 %	30.9 %	46.2 %	22.9 %	18.9 %
Paranday	727	76.9 %	29.2 %	47.7 %	23.1 %	11.1 %
Lucma	6,268	75.8 %	28.1 %	47.7 %	24.2 %	12.8 %
Julcán	13,356	75.6 %	29.2 %	46.4 %	24.4 %	17.4 %
Huaylillas	2,463	75.2 %	31.6 %	43.6 %	24.8 %	17.8 %
Huamachuco	55,281	75.0 %	32.3 %	42.7 %	25.0 %	22.5 %
Taurija	3,162	72.5 %	27.6 %	44.9 %	27.5 %	24.1 %
Santiago de Challas	2,946	72.3 %	24.4 %	47.9 %	27.7 %	21.5 %
Sayapullo	8,676	72.2 %	25.3 %	46.9 %	27.8 %	14.0 %
Urpay	3,180	72.0 %	27.1 %	44.9 %	28.0 %	23.6 %
Mache	3,372	70.6 %	23.6 %	47.0 %	29.4 %	11.6 %
Calamarca	6,616	70.1 %	25.3 %	44.8 %	29.9 %	17.8 %
Huaranchal	5,369	69.9 %	22.5 %	47.4 %	30.1 %	12.9 %
Buldibuyo	4,041	69.8 %	23.4 %	46.4 %	30.2 %	16.4 %
Huaso	6,593	69.2 %	22.9 %	46.3 %	30.8 %	17.6 %
Agallpampa	10,345	69.1 %	21.2 %	47.9 %	30.9 %	18.7 %
Bolívar	5,139	68.8 %	25.9 %	42.9 %	31.2 %	12.7 %
Charat	3,266	63.7 %	15.7 %	48.0 %	36.3 %	11.2 %
Compín	2,650	63.6 %	17.7 %	45.9 %	36.4 %	13.6 %
Angasmarca	6,299	63.4 %	18.8 %	44.6 %	36.6 %	15.5 %
Uchumarca	3,124	61.9 %	18.8 %	43.1 %	38.1 %	8.8 %
Parcoy	17,315	61.3 %	19.0 %	42.3 %	38.7 %	10.6 %
Santa Cruz de Chuca	3,360	59.2 %	14.4 %	44.8 %	40.8 %	14.7 %
Santiago de Chuco	20,671	58.0 %	16.4 %	41.6 %	42.0 %	11.9 %

Sigue en la página siguiente.

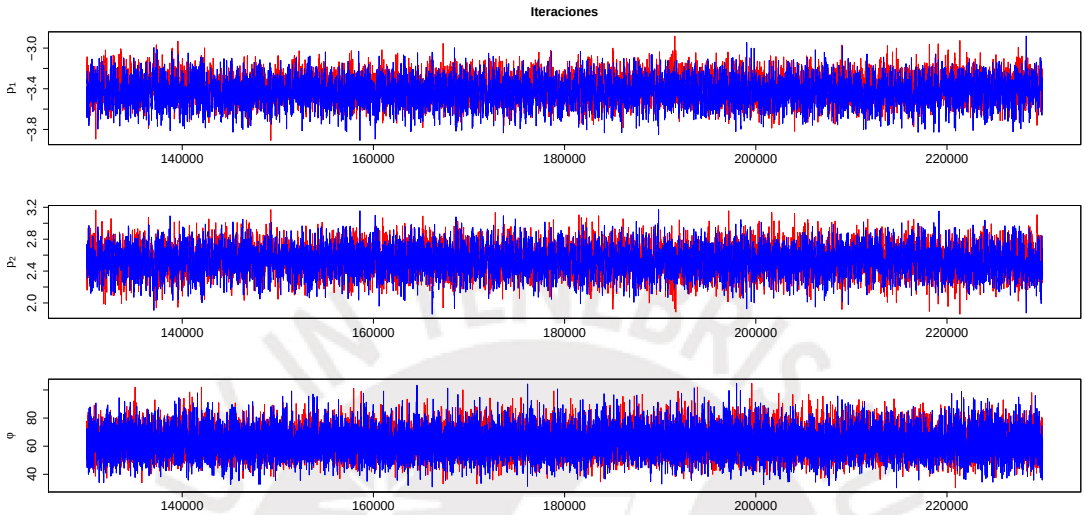
Distritos del departamento de La Libertad	Población	Pobres (%)			No Pobres (%)	Tasa de Analfabetismo (%)
		Total	Extremos	No extremos		
Longotea	2,494	57.7 %	16.0 %	41.7 %	42.3 %	9.4 %
Otuzco	26,664	56.4 %	15.0 %	41.4 %	43.6 %	12.2 %
Mollepata	2,860	55.5 %	12.8 %	42.7 %	44.5 %	21.9 %
La Cuesta	747	54.1 %	13.4 %	40.7 %	45.9 %	11.6 %
Cachicadán	6,935	53.3 %	14.3 %	39.0 %	46.7 %	15.0 %
Pacanga	18,809	53.2 %	10.2 %	43.0 %	46.8 %	10.4 %
Salpo	6,793	52.7 %	11.4 %	41.3 %	47.3 %	12.7 %
Quiruvilca	14,634	52.7 %	13.9 %	38.8 %	47.3 %	12.8 %
Pías	1,606	50.4 %	11.6 %	38.8 %	49.6 %	17.9 %
Mollebamba	2,035	49.7 %	10.7 %	39.0 %	50.3 %	15.7 %
Guadalupito	6,416	49.7 %	8.7 %	41.0 %	50.3 %	9.9 %
Chao	23,502	49.0 %	8.6 %	40.4 %	51.0 %	8.3 %
Virú	49,066	48.6 %	8.9 %	39.7 %	51.4 %	8.0 %
Poroto	3,650	48.5 %	9.1 %	39.4 %	51.5 %	11.4 %
Pueblo Nuevo	12,938	45.5 %	7.2 %	38.3 %	54.5 %	9.1 %
Cascas	15,404	44.1 %	8.0 %	36.1 %	55.9 %	10.7 %
Pataz	7,805	40.9 %	8.2 %	32.7 %	59.1 %	10.1 %
Paiján	23,913	39.9 %	5.9 %	34.0 %	60.1 %	6.8 %
Jequetepeque	3,548	39.9 %	6.4 %	33.5 %	60.1 %	7.5 %
San José	11,716	38.9 %	5.4 %	33.5 %	61.1 %	9.0 %
Guadalupe	38,225	38.5 %	6.1 %	32.4 %	61.5 %	7.7 %
Salaverry	14,080	38.0 %	5.1 %	33.0 %	62.0 %	3.3 %
Florencia de Mora	40,557	37.9 %	3.6 %	34.3 %	62.1 %	5.5 %
El Porvenir	142,413	33.9 %	4.2 %	29.7 %	66.1 %	5.6 %
Simbal	4,137	32.0 %	4.1 %	27.9 %	68.0 %	10.9 %
Ascope	7,229	31.4 %	3.9 %	27.5 %	68.6 %	6.4 %
Huanchaco	45,414	29.5 %	3.6 %	26.0 %	70.5 %	4.3 %
Rázuri	8,588	27.8 %	3.2 %	24.6 %	72.2 %	5.2 %
Chicama	15,523	27.5 %	3.0 %	24.5 %	72.5 %	6.4 %
Chepén	47,755	26.8 %	2.9 %	23.9 %	73.2 %	5.9 %
San Pedro de Lloc	16,576	25.7 %	2.6 %	23.1 %	74.3 %	4.6 %
La Esperanza	153,905	25.4 %	2.3 %	23.1 %	74.6 %	4.1 %
Moche	30,130	23.7 %	2.1 %	21.7 %	76.3 %	4.1 %
Pacasmayo	26,809	23.7 %	1.9 %	21.8 %	76.3 %	2.9 %
Laredo	33,270	21.3 %	2.3 %	18.9 %	78.7 %	5.9 %
Santiago de Cao	20,343	20.8 %	1.6 %	19.2 %	79.2 %	4.4 %
Casa Grande	30,813	19.5 %	1.5 %	18.0 %	80.5 %	4.7 %
Chocope	10,452	16.6 %	1.4 %	15.2 %	83.4 %	5.3 %
Magdalena de Cao	2,973	13.3 %	1.0 %	12.3 %	86.7 %	7.1 %
Víctor Larco Herrera	56,538	8.2 %	0.6 %	7.6 %	91.8 %	2.1 %
Trujillo	298,899	7.2 %	0.4 %	6.8 %	92.8 %	1.7 %

Tabla C.1: Distritos del departamento de La Libertad según condición de pobreza y tasa de analfabetismo, 2007. Fuente: Instituto Nacional de Estadística e Informática.

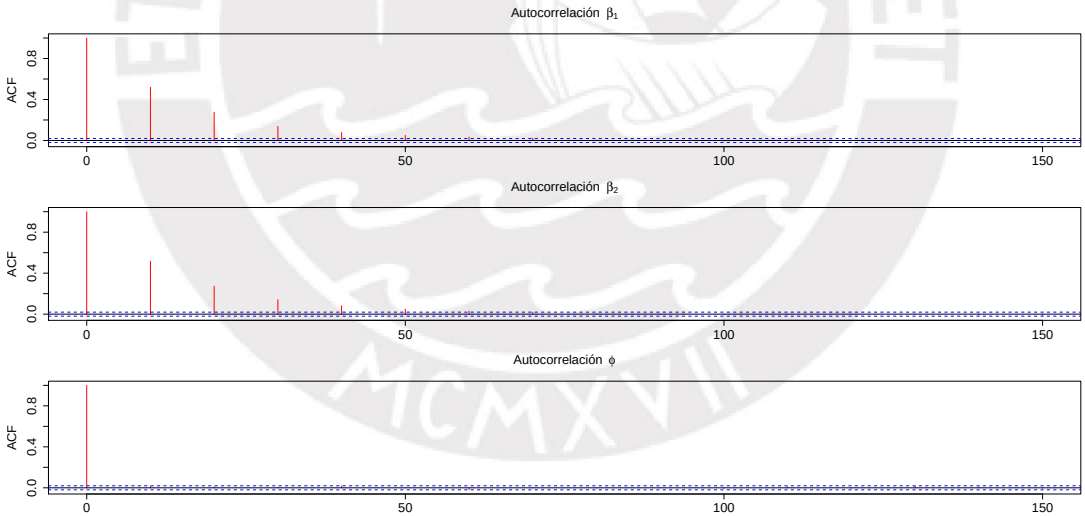
Apéndice D

Gráficos para los modelos de la aplicación

A continuación, se presentan los gráficos de valores simulados y de no existencia de autocorrelación para los modelos 1, 2, 3 y 4 desarrollados en el Capítulo 6.

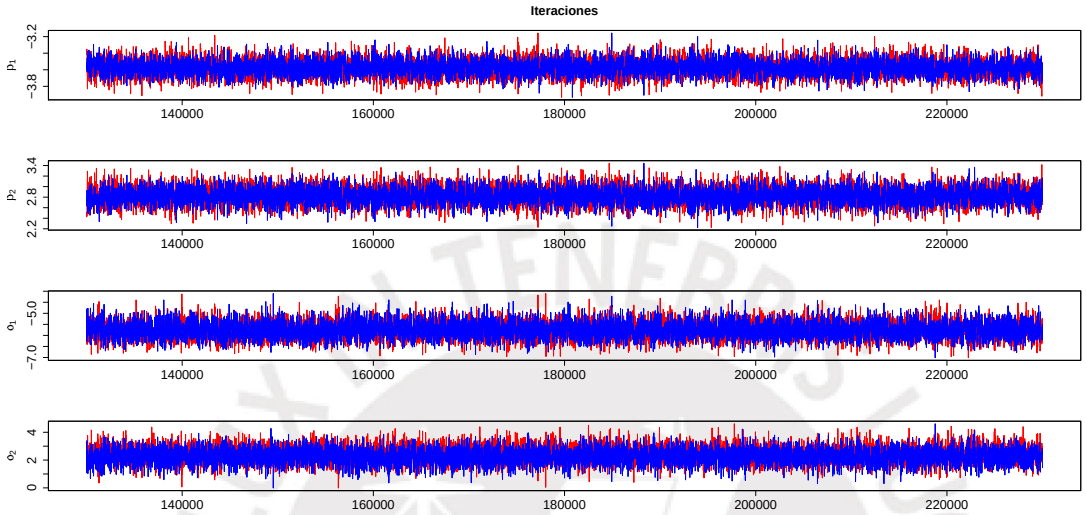


(a) Gráfico de los valores simulados de los parámetros cuando μ es variable y ϕ constante (modelo 1).

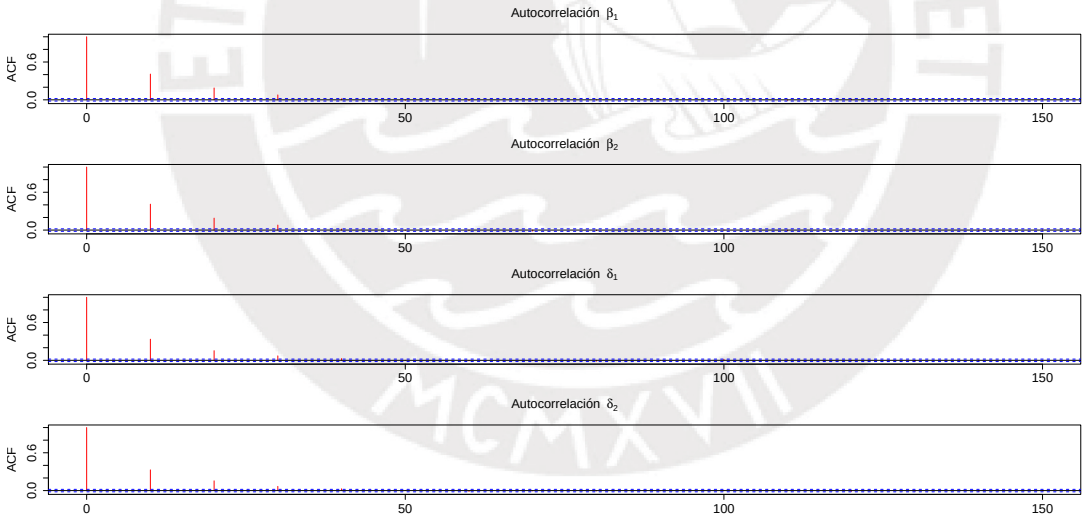


(b) Gráfico de no existencia de autocorrelación de los parámetros simulados cuando μ es variable y ϕ constante (modelo 1).

Figura D.1: Gráficos de valores simulados y de no existencia de autocorrelación para el modelo 1.

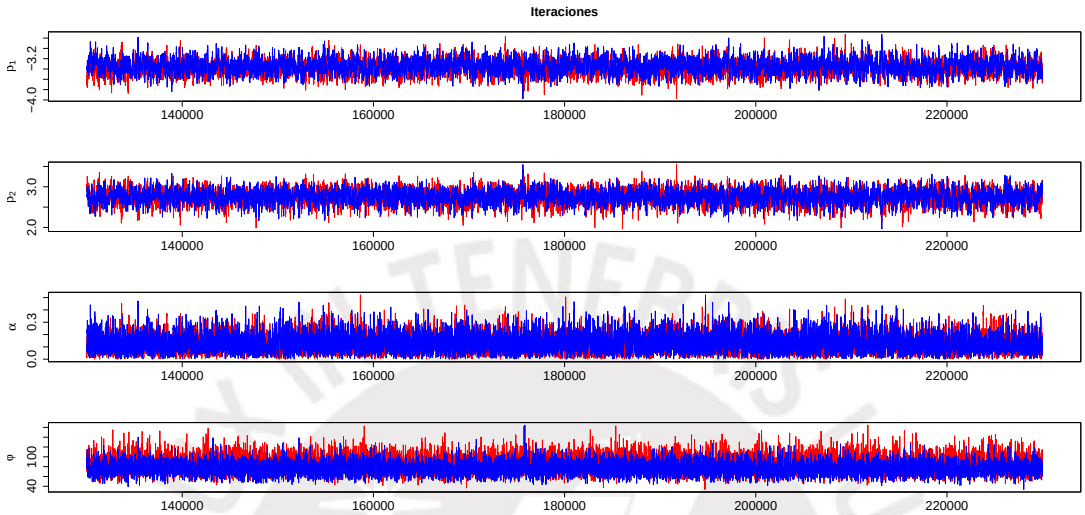


(a) Gráfico de los valores simulados de los parámetros cuando μ y ϕ son variables (modelo 2).

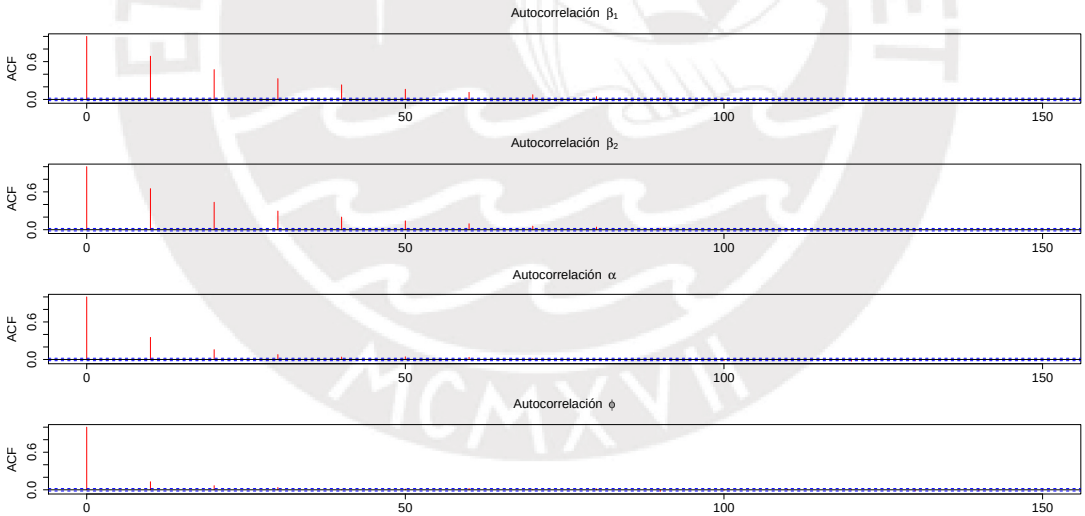


(b) Gráfico de no existencia de autocorrelación de los parámetros simulados cuando μ y ϕ son variables (modelo 2).

Figura D.2: Gráficos de valores simulados y de no existencia de autocorrelación para el modelo 2.

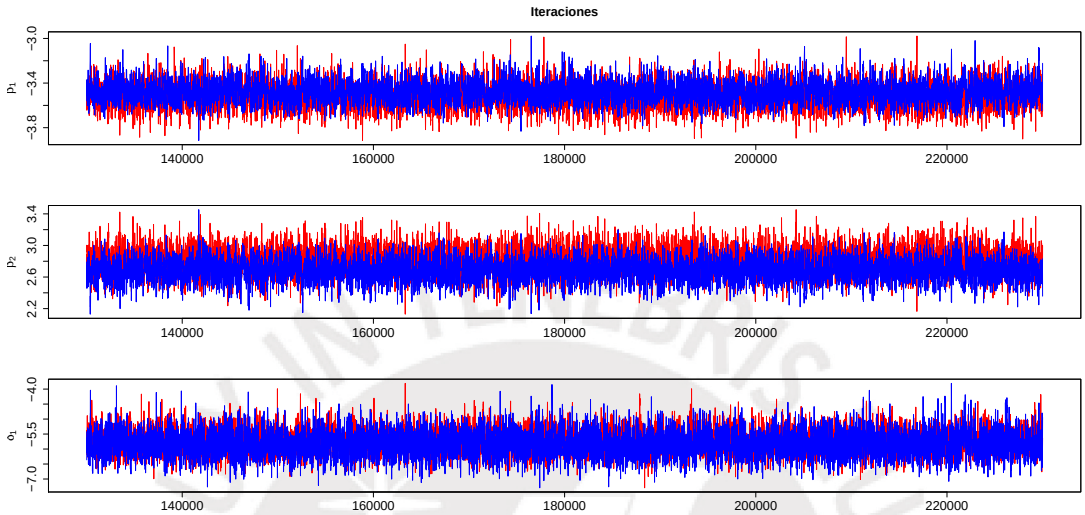


(a) Gráfico de los valores simulados de los parámetros cuando γ es variable y ϕ es constante (modelo 3).

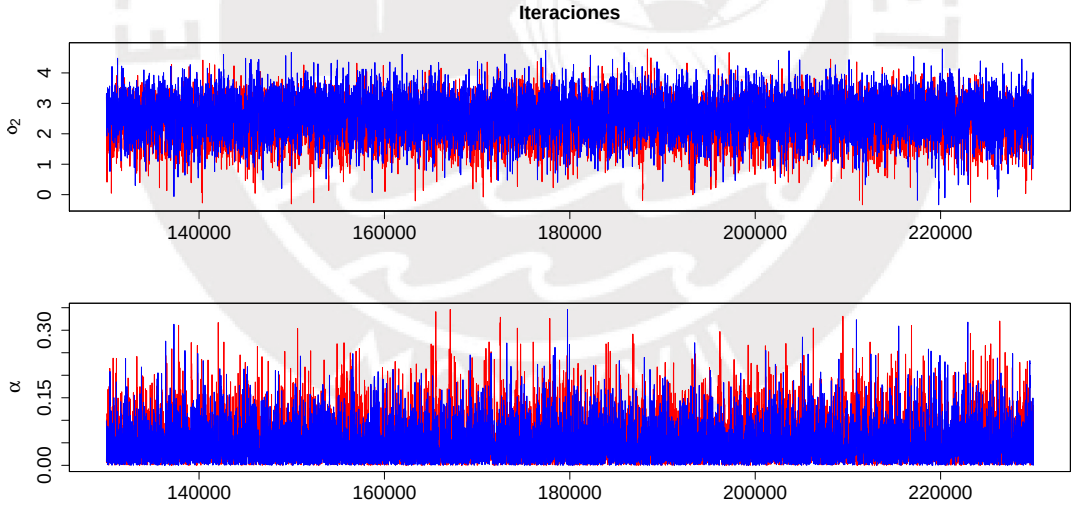


(b) Gráfico de no existencia de autocorrelación de los parámetros simulados cuando γ es variable y ϕ es constante (modelo 3).

Figura D.3: Gráficos de valores simulados y de no existencia de autocorrelación para el modelo 3.

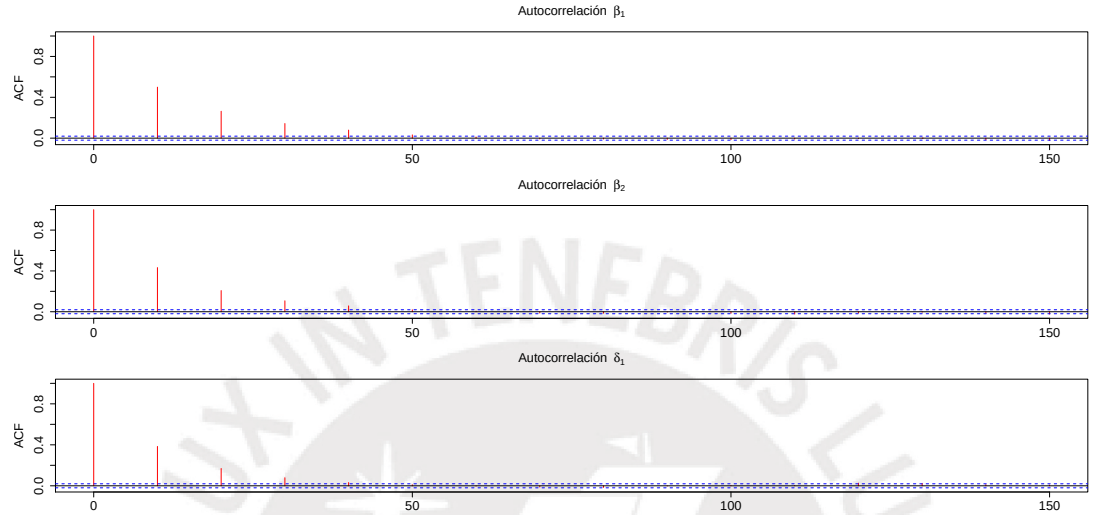


(a) Gráfico de los valores simulados de los parámetros cuando γ y ϕ son variables - Parte 1 (modelo 4).

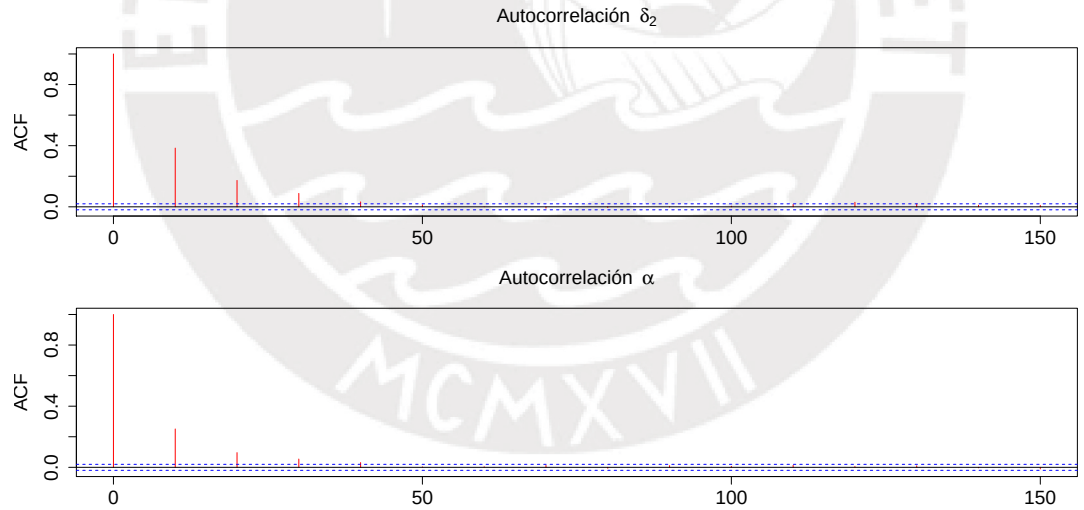


(b) Gráfico de los valores simulados de los parámetros cuando γ y ϕ son variables - Parte 2 (modelo 4).

Figura D.4: Gráficos de valores simulados para el modelo 4.



(a) Gráfico de no existencia de autocorrelación de los parámetros simulados cuando γ y ϕ son variables - Parte 1 (modelo 4).



(b) Gráfico de no existencia de autocorrelación de los parámetros simulados cuando γ y ϕ son variables - Parte 2 (modelo 4).

Figura D.5: Gráficos de no existencia de autocorrelación para el modelo 4.

Bibliografía

Alencar, F. H. C. d. (2016). *Diagnóstico de influência para uma família de modelos de regressão para dados de taxas e proporções*, Tesis de maestría, Universidade Federal de Pernambuco.

Bayes, C. L., Bazán, J. L. y García, C. (2012). A New Robust Regression Model for Proportions, *Bayesian Analysis* **7(4)**: 841–866.

Bouguila, N., Monga, E. y Ziou, D. (2006). Practical Bayesian estimation of a finite beta mixture through gibbs sampling and its applications, *Statistics and Computing* **16(2)**: 215–225.

Brooks, S. P. (2002). Discussion on the paper by Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A., *Journal of the Royal Statistical Society, Series B* **64(4)**: 616–639.

Chia, L. (2012). *Análisis Clásico y Bayesiano en el Modelo de Distribución Beta Rectangular*, Tesis de maestría, Pontificia Universidad Católica del Perú.

Cribari-Neto, F. y Zeileis, A. (2010). Beta Regression in R , *Journal of Statistical Software* **34(2)**: 1–24.

Ferrari, S. y Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions , *Journal of Applied Statistics* **31(7)**: 799–815.

Gentle, J. (2002). *Elements of Computational Statistics*, Springer, Verlag, New York.

Hahn, E. (2008). Mixture densities for project management activity times: A robust approach to PERT, *European Journal of Operational Research* **188(2)**: 450–459.

INEI (2015). Evolución de la Pobreza Monetaria, 2009-2014, *Informe técnico*, Instituto Nacional de Estadística e Informática.

Krishnamoorthy, K. (2006). *Handbook of Statistical Distributions with Applications*, Chapman & Hall/CRC.

Markatou, M. (2000). Mixture Models, Robustness, and the Weighted Likelihood Methodology, *Biometrics* **56(2)**: 483–486.

Mira, A. (2005). MCMC Methods to Estimate Bayesian Parametric Models, *Handbook of Statistics* **25**: 415–436.

Paolino, P. (2001). Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables, *Political Analysis* **9(4)**: 325–346.

Simas, A., Barreto-Souza, W. y Rocha, A. (2010). Improved estimators for a general class of beta regression models, *Computational Statistics and Data Analysis* **54(2)**: 348–366.

Smithson, M. y Verkuilen, J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables, *Psychological Methods* **11(1)**: 54–71.

Spiegelhalter, D. J., Best, N.and Carlin, B. y Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society: Series B* **64(4)**: 583–639.