UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2017, Vol. 10, No. 1, pp. 41–56

Análisis de distribuciones a priori para los parámetros de escala del modelo *ZIP*

Analysis of prior distributions for the scales parameters of the ZIP model

Juan Daniel Molina Muñoz[a]
jdmolinam@unal.edu.co

Isabel Cristina Ramírez Guevara[b]
iscramirezgu@unal.edu.co

Resumen

En el presente artículo se plantea la evaluación de un conjunto de distribuciones a priori para los parámetros de escala del modelo de regresión Poisson inflado con ceros (*Zero Inflated Poisson, ZIP*). Tradicionalmente, se utiliza la distribución Gamma inversa como a priori para los parámetros de escala. Algunos estudios han mostrado que cuando los valores de los hiperparámetros de esta distribución son muy pequeños, las inferencias a posteriori no son adecuadas. El interés se centra en evaluar tres distribuciones a priori para los parámetros de escala del modelo: la Gamma inversa; la Half Cauchy, que se ha usado para la situación planteada y que ha demostrado funcionar adecuadamente;la beta 2 escalada (SBeta2), la cual es una distribución de colas pesadas que tiene un comportamiento mejor en el origen y en la cola derecha.

Se desarrolla un estudio de simulación con el que se pretende analizar el efecto de la distribución a priori asignada a los parámetros de escala sobre el encogimiento en las estimaciones a posteriori de los parámetros, y se evalúa ante la presencia de observaciones atípicas cómo el ajuste que el modelo realiza con cada una de las distribuciones a priori candidatas para los parámetros de escala. El análisis se centra en estas dos características (encogimiento en las estimaciones a posteriori de los parámetros y ajuste de observaciones atípicas), pues son estas las principales críticas que diferentes autores plantean al uso de la distribución Gamma inversa como a priori para los parámetros de escala. Finalmente, se presenta una aplicación con datos reales.

***Palabras clave*:** Inferencia bayesiana, modelo *ZIP*, parámetros de escala, distribución SBeta2, distribución Half Cauchy, distribución Gamma inversa.

[a]M.Sc Estadística, Escuela de Estadística - Universidad Nacional de Colombia, sede Medellín, Colombia

[b]Profesor, Escuela de Estadística - Universidad Nacional de Colombia, sede Medellín, Colombia

Abstract

In this paper, it is proposed the evaluation of a set of prior distributions for the scale parameters of the Zero-Inflated Poisson Regression model (*ZIP*). Traditionally the inverse-gamma distribution is used as prior for scale parameters. Some studies have shown that when the values of the hyperparameters of this distribution are very small, inferences are not adequate. We focus on evaluating three prior distributions for modeling scale parameters: inverse-gamma; half Cauchy and scaled beta 2 (SBeta2). The half Cauchy has been used in the situation in question and has proven to work properly. The SBeta2 is a heavy-tailed distribution that has better performance at the origin and at the right tail.

A simulation study is developed, with which we intend to analyze the effect of the prior distribution assigned to the scale parameters on the shrinkage of the posterior estimates of parameters. Besides, the presence of outliers is evaluated regarding the adjustment of the corresponden values. This is done for each of the three prior distributions considering. The analysis focuses shrinkage of the posterior estimates of parameters and adjustment of outliers because the main criticisms on the use of the inverse-gamma distribution concentrate on this two issues. Finally an application is presented with real data.

***Keywords*:** Bayesian inference, ZIP model, scales parameters, SBeta2 distribution, Half Cauchy distribution, Inverted-gamma distribution.

1. Introducción

Para el modelamiento de fenómenos de conteo con presencia excesiva de ceros deben considerarse modelos especiales que se ajustan a dicha condición. Uno de los modelos más utilizados en este contexto es el modelo *ZIP* propuesto por Lambert (1992). Ghosh et al. (2006) plantean la opción de aplicar dicho modelo desde el enfoque bayesiano, buscando así un mejor comportamiento cuando se tienen muestras pequeñas, o una proporción muy grande de ceros respecto al total de datos.

Dentro del enfoque bayesiano, una de las decisiones fundamentales es la determinación de la distribución a priori de los parámetros de un modelo. En este caso, el interés se centra en evaluar el impacto de la distribución a priori para los parámetros de escala del modelo *ZIP*. Con este fin se estudian tres distribuciones: 1. la Gamma inversa, la cual ha sido ampliamente utilizada como a priori para los parámetros de escala en modelos jerárquicos; sin embargo, diferentes autores han planteado fuertes críticas a esta práctica, como Berger (2006), quien plantea que el uso de dicha distribución como a priori para la varianza conduce a una distribución a posteriori sesgada en valores cercanos a cero, esto puede llevar, a su vez, a resultados incoherentes y a la incapacidad de predecir o ajustar observaciones atípicas. Gelman (2006), por su parte, argumenta que la Gamma inversa(ϵ, ϵ) cuando se usa como a priori para la varianza y buscando que sea no informativa se hace $\epsilon \to 0$, lo

cual produce en realidad produce un encogimiento en las estimaciones a posteriori de los parámetros del modelo y, si por la naturaleza de los datos fuera posible, también produciría valores pequeños de la varianza, así la a priori se conviertiría en informativa. El autor además ilustra por medio de un ejemplo con datos reales el problema de concentración alrededor del cero.

La segunda alternativa para evaluar es la distribución half-Cauchy, la cual es estudiada por Gelman (2006) como a priori para la desviación estándar en modelos jerárquicos, mostrando que se comporta adecuadamente. Y por último, la tercera alternativa es la distribución Beta2 escalada (SBeta2) propuesta para este uso por Pericchi (2010), la cual, posee unas propiedades teóricas convenientes cuando se usa como a priori para parámetros de escala.

Con el fn de evaluar el impacto de la distribución a priori asignada a los parámetros de escala en el modelo *ZIP* se realizó un estudio de simulación, en el que cada distribución a priori candidata fue analizada bajo condiciones de encogimiento en las estimaciones a posteriori de los parámetros y la capacidad del modelo de ajustar observaciones atípicas.

El análisis se centra en estas dos características pues son las principales críticas que diferentes autores, como Berger (2006) y Gelman (2006), plantean al uso de la distribución Gamma inversa como a priori para los parámetros de escala.

Las siguientes secciones del presente artículo están organizadas de a siguiente manera: en la sección 2 se presenta la definición del modelo *ZIP* y algunas propuestas de distribuciones a priori para parámetros de escala. En la sección 3 se desarrolla el estudio de simulación, se realiza una definición de las características generales y condiciones del mismo, se presentan y analizan sus resultados. En la sección 4 se presenta una aplicación con datos reales. Finalmente, en la sección 5 se presentan las principales conclusiones obtenidas de este trabajo.

2. Modelo *ZIP*

Las variables que representan fenómenos de conteo deben modelarse a través de distribuciones discretas, como la distribución Poisson. Sin embargo, existen casos en los que el número de ceros que presenta la variable estudiada supera la frecuencia teórica que se espera según la distribución definida a su ajuste. En estos casos se manifiesta que los datos presentan un exceso de ceros o que están "inflados con ceros". Si se presenta un exceso de ceros es un error pensar que los datos se ajustan a una distribución discreta tradicional, pues cualquier inferencia realizada bajo esta idea sería incorrecta (Heibron 1994), por lo cual se hace necesario usar un modelo inflado con ceros.

El modelo *ZIP* parte del modelo de regresión Poisson clásico, y consiste en la combinación lineal de distribuciones de probabilidad. Este modelo es utilizado con frecuencia para trabajar datos de conteo con exceso de ceros, el cual fue propuesto por Lambert (1992). Bajo este modelo se tienen dos clases de ceros: los generados

por la distribución Poisson que aparecen con probabilidad $1-p$, y un conjunto de ceros extra que aparecen con probabilidad p.

Siendo $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ el vector de la variable respuesta de la regresión, bajo el modelo *ZIP* las Y_i tienen la siguiente probabilidad:

$$P(Y_i = y) = \begin{cases} p_i + (1-p_i)\exp(-\lambda_i) & \text{para} \quad y = 0 \\ (1-p_i)\dfrac{\exp(-\lambda_i)\lambda_i^y}{y!} & \text{para} \quad y = 1, 2, \ldots \end{cases}$$

Lo anterior se denota como $Y_i \sim ZIP(p_i, \lambda_i)$. Se asume entonces que la variable respuesta está relacionada con las covariables de la regresión a partir de la estructura de los modelos lineales generalizados, en función de $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$ y $\mathbf{p} = (p_1, \ldots, p_n)^T$, de la siguiente forma:

$$\begin{aligned} \log(\boldsymbol{\lambda}) &= (\log(\lambda_1), \ldots, \log(\lambda_n))^T = \mathbf{B}\boldsymbol{\beta}, \\ \text{logit}(\mathbf{p}) &= (\text{logit}(p_1), \ldots, \text{logit}(p_n))^T = \mathbf{G}\boldsymbol{\gamma}, \end{aligned}$$

donde $\text{logit}(p_i) = \log\left(\dfrac{p_i}{1-p_i}\right)$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ y $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_k)^T$ son los vectores que contienen los parámetros del modelo, con k igual al número de covariables; $\mathbf{B}$ y $\mathbf{G}$ son matrices conocidas en función de las covariables de la regresión, cada una con dimensión $(n, k+1)$.

2.1. Propuestas de distribuciones a priori para parámetros de escala

Gelman (2006) presenta un conjunto de distribuciones a priori no informativas para los parámetros de escala de los modelos jerárquicos. Se plantea una nueva familia de distribuciones a priori condicionadas conjugadas, denominada *folded-noncentral-t*, para los parámetros de la desviación estándar. Por medio de un ejemplo se ilustran los serios problemas que puede presentar la familia Gamma inversa de distribuciones a priori no informativas, de esta forma se cuestiona el uso tan frecuente de esta distribución como a priori para la varianza de un modelo. Tambiés, se estudia el uso de la distribución half-Cauchy, la cual pertenece a la familia half-t, como a priori para la desviación estándar en modelos jerárquicos, mostrando que se comporta adecuadamente, pues asintóticamente es una a priori no informativa, mientras que para los valores suficientemente grandes de su hiperparámetro es una a priori débilmente informativa. Hay que tener en cuenta que al mismo tiempo es una distribución flexible y presenta un buen comportamiento alrededor del cero.

Por otro lado, Fúquene et al. (2014) proponen una nueva clase de distribuciones a priori hipergeométricas de colas anchas, que resulta de la combinación de la distribución t-student para el parámetro de localización y la distribución beta2 escalada (SBeta2) para el cuadrado del parámetro de escala. De estas distribuciones a priori pueden obtenerse colas más pesadas que las de distribuciones a priori t-student y

la varianza puede presentar un comportamiento más adecuado respecto al origen y a las colas.

Pérez et al. (2016) proponen la distribución SBeta2 como alternativa para las a priori de la varianza y la precisión, en lugar de la distribución Gamma inversa. Entre las ventajas de la SBeta2 están: si la varianza distribuye SBeta2, entonces la precisión también, lo que se conoce como "propiedad de reciprocidad". Es posible simular valores de la SBeta2, y la distribución puede integrarse al esquema del muestreador de Gibbs. Es una distribución flexible, dado que se pueden modelar diferentes tipos de comportamientos en el origen y la cola. La SBeta2 se rige por tres parámetros que son factibles de elicitar, uno rige el comportamiento en el origen, otro el de la cola derecha y el tercero la escala de la distribución. Finalmente, la SBeta2 es una distribución robusta, donde el espesor de su cola es equivalente al de la t-student (Pérez et al. 2016).

3. Estudio de simulación

La comparación de las distribuciones a priori para los parámetros de escala del modelo *ZIP* se realizó vía simulación. Se partió del caso más simple del modelo *ZIP*, en que se consideró una única variable regresora y sólo se consideró intercepto para la ecuación asociada con la proporción extra de ceros. De esta forma, el modelo se resume en la siguiente expresión:

$$\begin{aligned} &Y_i \sim \text{ZIP}(p_i, \lambda_i), \\ &\log(\lambda_i) = \beta X, \\ &\text{logit}(p_i) = \gamma_0 + \gamma X. \end{aligned}$$

Se asume que $\beta \sim N(0, \sigma_1^2)$, $\gamma \sim N(0, \sigma_2^2)$ y $\gamma_0 \sim U(-2.5, 2.5)$. Durante el estudio de simulación se asumió $X \sim U(0, 1)$ además σ_1^2 y σ_2^2 fueron valores fijos. Se desarrollaron en total 48 escenarios, compuestos por las siguientes condiciones: 3 distribuciones candidatas como a priori para los parámetros de escala del modelo *ZIP*: Gamma inversa, half-Cauchy y SBeta2; 4 valores asignados a los parámetros de escala: $\sigma_1^2 = \sigma_2^2 = 0.1, 3, 10, 35$; 4 tamaños muestrales: $n = 5, 15, 30, 100$. Cada uno de los escenarios se simuló 1000 veces.

A continuación se enlista el conjunto de pasos que se llevaron a cabo en el desarrollo del estudio de simulación:

1. Para una determinada distribución candidata, se construyó el código del modelo *ZIP*, ajustando la distribución candidata como a priori para sus parámetros de escala.

2. Para un determinado escenario de simulación, se generaron los datos de la variable respuesta que distribuye *ZIP* y de la covariable. Para generar los valores de la variable respuesta se parte de la estructura del modelo considerada.

3. Para un determinado escenario, se simularon las cadenas a posteriori de los parámetros del modelo *ZIP*, esto se hizo por medio del método MCMC (Markov Chain Monte Carlo), a través del software `OpenBUGS`, el cual usa como insumos el modelo *ZIP* ajustado con una determinada distribución candidata y los datos generados de la variable respuesta y de la covariable.

Las distribuciones candidatas como a priori para los parámetros de escala del modelo *ZIP* se trabajaron bajo las siguientes condiciones: Gamma inversa$(0.01, 0.01)$, pues tradicionalmente cuando se usa esta distribución como a priori para parámetros de escala se escogen hiperparámetros pequeños (Gelman 2006), con el próposito de obtener una a priori no informativa. Gelman (2006) utiliza la half-Cauchy(25) como a priori para la desviación estándar; sin embargo, a partir de la relación entre la distribución Beta2 y la half-Cauchy en la que Polson & Scott (2012) demuestran que existe, Pérez et al. (2016) muestran que usar la half-Cauchy(25) como a priori para la desviación estándar es equivalente a usar la SBeta2$(0.5, 0.5, 25^2)$ como a priori para la varianza. Finalmente, Pérez et al. (2016) muestran que la SBeta2$(1, 1, 25^2)$ presenta un comportamiento adecuado cuando es usada como a priori para la varianza.

Se realizó una comparación sobre las distribuciones candidatas en términos del encogimiento de los parámetros principales del modelo jerárquico: β y γ, procediendo así, de forma similar a la metodología planteada por Fruhwirth-Schnatter & Wagner (2010). En cada uno de los escenarios de la simulación se calculó el RMSE (Raíz del error cuadrático medio), donde por ejemplo para la estimación del parámetro β el RMSE se calcula de la siguiente forma:

$$\text{RMSE} = \sqrt{\sum_{i=1}^{1000} \frac{(\beta - \hat{\beta}_i)^2}{1000}}$$

donde, para un determinado escenario, en cada una de las 1000 simulaciones del mismo, $\hat{\beta}_i$ se calcula como la mediana de la cadena a posteriori del parámetro β.

Entre mayor sea el RMSE, implica que es más grande el problema de encogimiento en las estimaciones a posteriori de los parámetros. Se presentan resultados para cuatro condiciones principales: fijando los parámetros de escala en $\sigma_1^2 = \sigma_2^2 = 0.1, 3, 10, 35$.

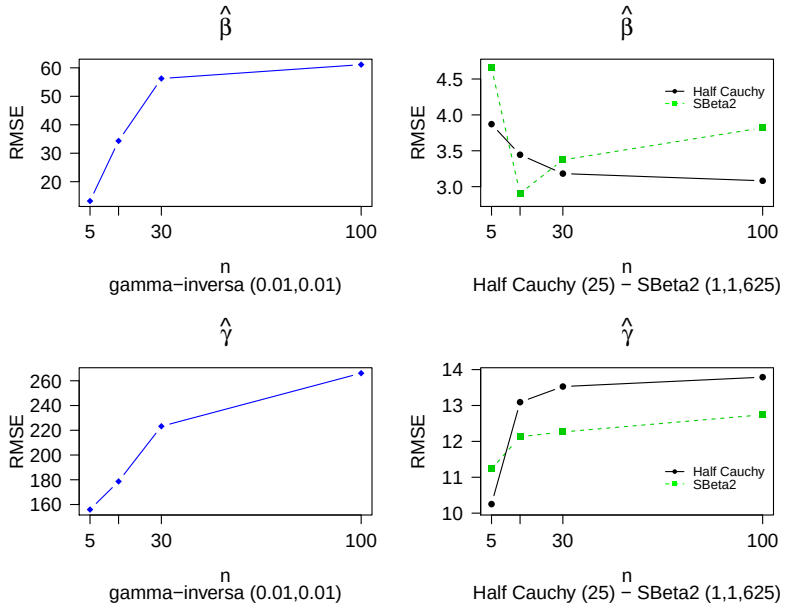

Figura 1: *RMSE VS Tamaño muestral* - $\sigma_1^2 = \sigma_2^2 = 0.1$. *Fuente: elaboración propia.*

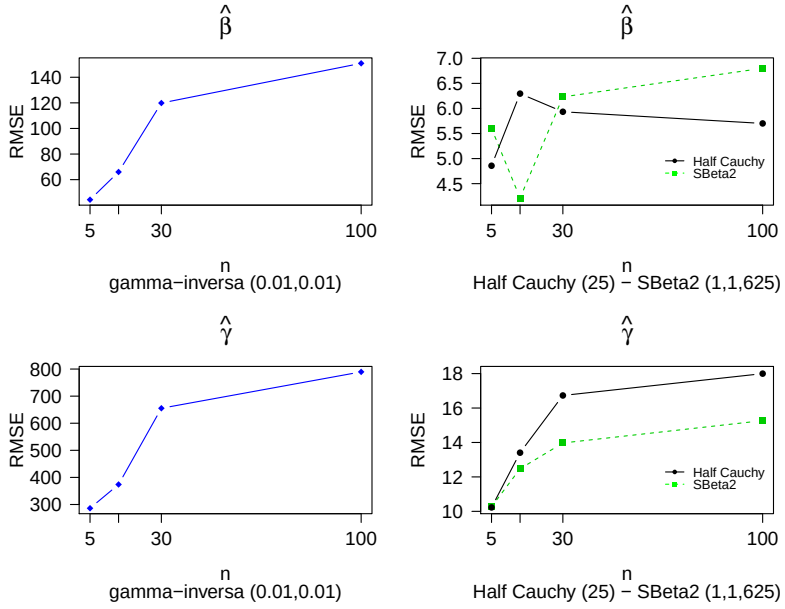

Figura 2: *RMSE VS Tamaño muestral* - $\sigma_1^2 = \sigma_2^2 = 3$. *Fuente: elaboración propia.*

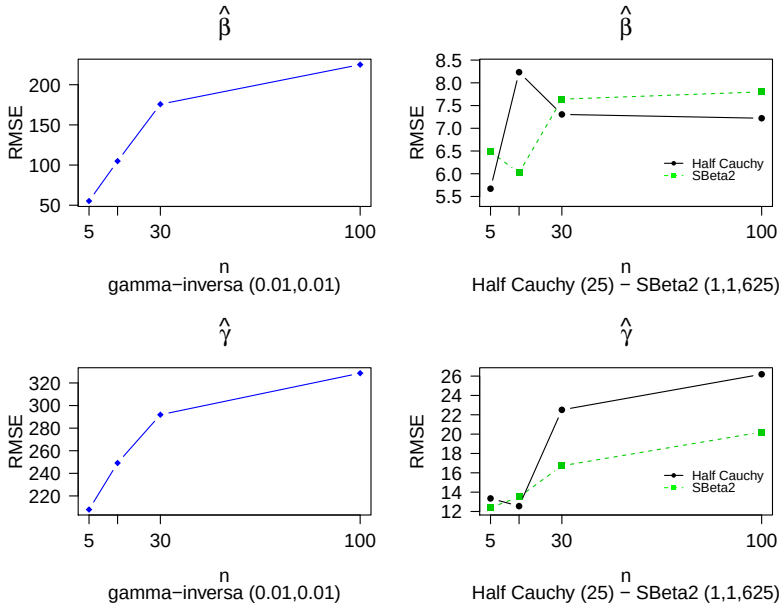

Figura 3: *RMSE VS Tamaño muestral* - $\sigma_1^2 = \sigma_2^2 = 10$. *Fuente: elaboración propia.*

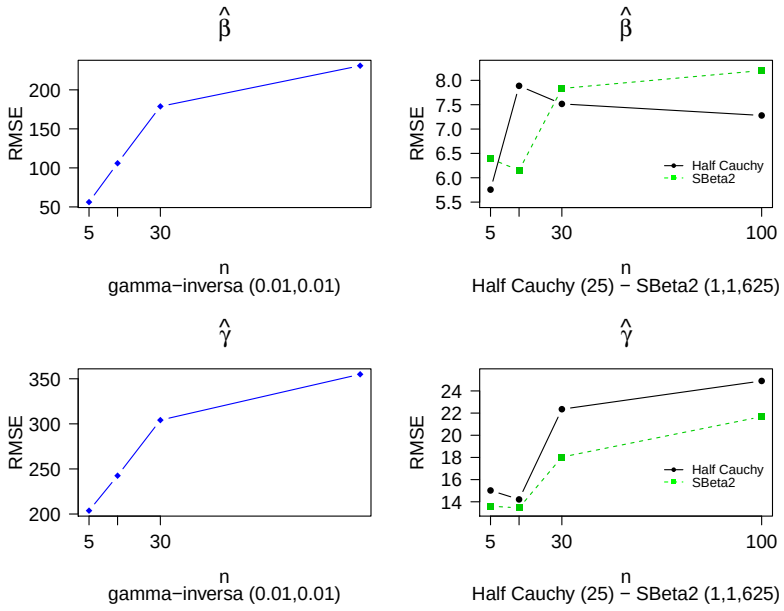

Figura 4: *RMSE VS Tamaño muestral* - $\sigma_1^2 = \sigma_2^2 = 35$. *Fuente: elaboración propia.*

En general, en cada una de las gráficas del análisis de encogimiento (figuras 1, 2, 3 y 4) tanto para $\hat{\beta}$ como para $\hat{\gamma}$ el RMSE se reduce considerablemente con las distribuciones half-Cauchy y la SBeta2 en comparación con la Gamma inversa, los resultados de la half-Cauchy y la SBeta2 son relativamente similares. En algunas gráficas se observa que el RMSE aumenta con el tamaño muestral, o una alternación entre crecimiento-decrecimiento, todo esto puede explicarse como una resolución de conflicto; es decir, el impacto de la a priori se reduce cuando aumenta el tamaño muestral (O'Hagan & Pericchi 2012).

3.1. Chequeo de convergencia

El método MCMC utilizado para la obtención de las cadenas a posteriori de los parámetros del modelo está basado en el supuesto de que las cadenas alcanzan la distribución estacionaria. Por esto resulta necesario hacer un chequeo de convergencia sobre las cadenas a posteriori obtenidas en este estudio de simulación. En el presente trabajo, el chequeo de convergencia se lleva a cabo de forma similar al procedimiento planteado para dicho fin por Barrera & Correa (2008). Aquí se entiende que para una determinada cadena, el chequeo consiste en evaluar la autocorrelación existente entre los valores generados del parámetro en distintos rezagos, realizar un gráfico de promedios móviles y por último se realizar un test para verificar la convergencia de la cadena. El test utilizado es el KPSS (Kwiatkowski-Phillips-Schmidt-Shin), con el cual se evalúa el siguiente conjunto de hipótesis:

$$\begin{gathered} H_0 = \text{La cadena ha alcanzado la distribución estacionaria} \\ VS \\ H_1 = \text{La cadena no ha alcanzado la distribución estacionaria} \end{gathered}$$

Para tomar una decisión sobre la prueba de hipótesis, el test KPSS se basa en el estadístico de prueba LM, el cual fue desarrollado por Kwiatkowski et al. (1992).

A continuación de forma ilustrativa se presentan los resultados del chequeo de convergencia para la cadena obtenida bajo las condiciones: $\sigma_1^2 = \sigma_2^2 = 0.1$, $n = 15$, distribución candidata SBeta2, simulación número 83 del parámetro β. La tabla 1 presenta los valores de la autocorrelación entre los valores generados del parámetro con diferentes rezagos, de dichos resultados se observa que los valores de autocorrelación están muy cerca del cero, con lo cual se descarta la existencia de una relación lineal entre los elementos de la cadena.

Tabla 1: *Autocorrelación - Cadena del análisis de encogimiento. Fuente: elaboración propia.*

	1 rezago	5 rezagos	10 rezagos	50 rezagos
β	−0.017005537	−0.001269910	−0.003609066	0.001738451

La figura 5 presenta los promedios móviles de los valores generados del parámetro. Del gráfico se observa una pronta estabilización de dichos promedios. Finalmente,

por medio del software estadístico R, se realiza el test KPSS para el cual se obtiene que el valor del estadístico de prueba es 0.0269 y valor p de 0.1, con lo cual se concluye que no existe suficiente evidencia muestral para rechazar la hipótesis nula. Así, dados los resultados de autocorrelación, del gráfico de promedio móviles y el test KPSS, se concluye que la cadena a posteriori, bajo las condiciones establecidas, alcanza la distribución estacionaria. Es de mencionar que todas las demás cadenas del estudio de simulación cumplen con el supuesto de alcanzar la distribución estacionaria.

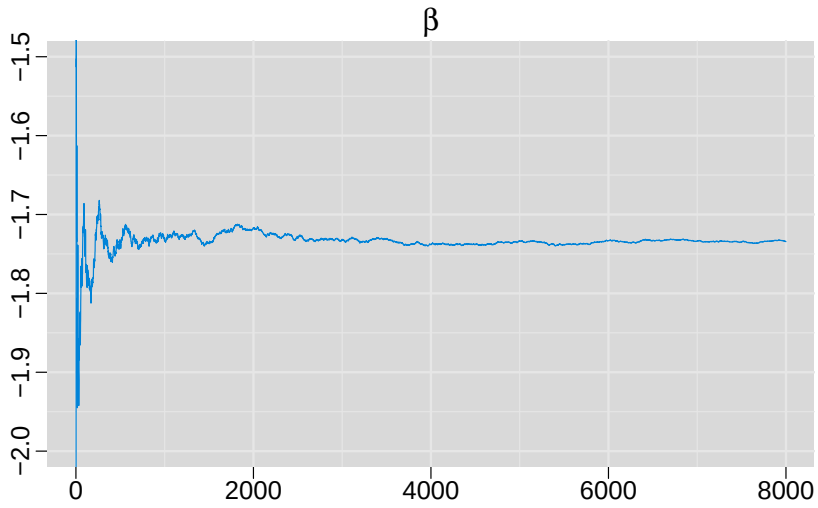

Figura 5: *Promedios móviles - Cadena del análisis de encogimiento. Fuente: elaboración propia.*

3.2. Análisis de la capacidad del modelo de ajustar observaciones atípicas

Otra circunstancia bajo la cual se evalúan las distribuciones candidatas como a priori para los parámetros de escala del modelo *ZIP* es el análisis del ajuste que el modelo realiza de observaciones atípicas con cada candidata. Dicha circunstancia es evaluada dada la problemática que algunos autores plantean sobre la Gamma inversa cuando se usad como a priori para los parámetros de escala, en cuanto al ajuste inadecuado de observaciones atípicas, esto debido a que con la Gamma inversa las predicciones y ajustes se centran en la media a posteriori (Gelman 2006), (Berger 2006).

Para el análisis del ajuste que el modelo *ZIP* realiza de observaciones atípicas, se utilizaron los datos ya generados en el análisis de encogimiento, fijando los parámetros de escala $\sigma_1^2 = \sigma_2^2 = 0.1$, haciendo comparaciones para las diferentes distribuciones candidatas y los diferentes tamaños muestrales. Así, para una de-

terminada candidata y un determinado tamaño muestral, se tomaron los datos de la variable respuesta y a estos se les agregó deliberadamente una observación atípica, donde el valor atípico se definió como dos veces el máximo valor de los datos originales de la variable respuesta del modelo.

El valor esperado de la variable respuesta del modelo se definió como el ajuste de la observación atípica, teniendo en cuenta que si $Y \sim \text{ZIP}(p, \lambda)$, entonces $E(Y) = (1-p)\lambda$. De esta forma, para una determinada distribución candidata, para un determinado tamaño muestral, nuevamente se generaron las cadenas a posteriori de los parámetros del modelo, cambiando los datos originales por los contaminados. Finalmente, a partir de las cadenas a posteriori se calculaba el valor esperado de la variable respuesta del modelo, el cual, como ya se mencionó, se definió como el ajuste que el modelo ofrece de las observaciones atípicas. A continuación se presentan los resultados del RMSE del ajuste de las observaciones atípicas respecto al verdadero valor de las mismas.

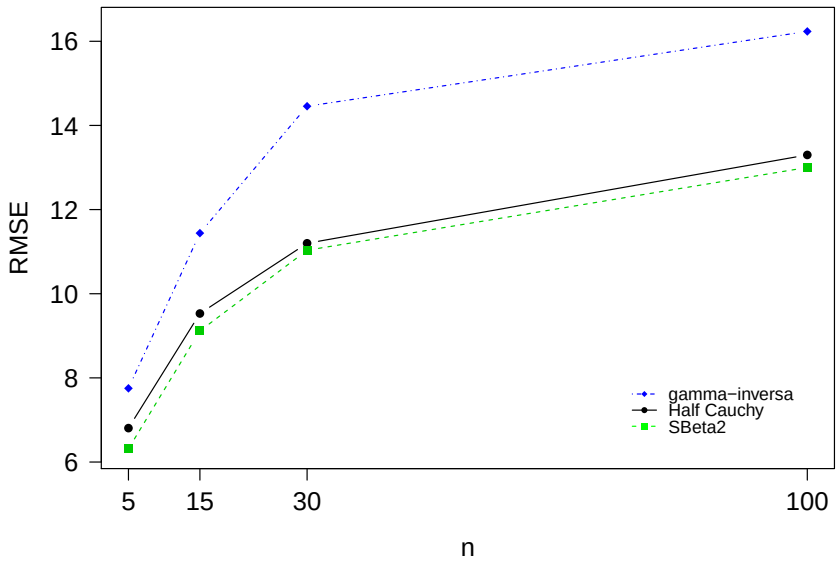

Figura 6: *RMSE VS Tamaño muestral - Ajuste de una observación atípica. Fuente: elaboración propia.*

De la figura 6 se observa que los resultados para las tres distribuciones candidatas es relativamente similar en cuanto que en todas se observa un aumento del RMSE con el tamaño muestral, esto nuevamente puede explicarse como una resolución del conflicto entre la a priori y los datos (O'Hagan & Pericchi 2012). Sin embargo, para cualquier tamaño muestral la distribución Gamma inversa presenta mayor error en el ajuste de la observación atípica. Los resultados para las distribuciones

half-Cauchy y la SBeta2 son relativamente similares, aunque los errores en el ajuste de la observación atípica con la SBeta2 siempre son menores.

4. Caso práctico

Se presenta una aplicación con datos de cultivo de manzanas, obtenidos por Marin et al. (1993). Los datos son el número de raíces producidas por 270 brotes micropropagados de la columna de cultivos de manzana tipo Trajan. Los brotes crecieron en medios que contenían diferentes concentraciones de proteína BAP y en cámaras de cultivo expuestas a condiciones de fotoperiodo de 8 y 16 horas. Así, estos datos conforman un modelo de regresión, donde la variable respuesta es el número de raíces en los brotes, y las covariables son la concentración de la proteína BAP en el medio y la condición de fotoperiodo. Rodrigues (2006) mostró que la variable respuesta de estos datos distribuye *ZIP*, de esta forma es pertinente trabajar bajo el modelo *ZIP*, con la siguiente estructura:

$$\begin{aligned} &Y_i \sim ZIP(p_i, \lambda_i), \\ &\log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \\ &\text{logit}(p_i) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2. \end{aligned}$$

donde Y representa el número de raíces en cada uno de los brotes, X_1 representa la condición de fotoperiodo y X_2 representa la concentración de la proteína BAP. Se asume que $\beta_0 \sim U(-2.5, 2.5)$, $\beta_1 \sim N(0, \sigma_1^2)$, $\beta_2 \sim N(0, \sigma_2^2)$, $\gamma_0 \sim U(-2.5, 2.5)$, $\gamma_1 \sim N(0, \sigma_3^2)$, $\gamma_2 \sim N(0, \sigma_4^2)$.

En este caso práctico, el análisis de las distribuciones a priori de los parámetros de escala del modelo *ZIP* se realiza a partir de tres condiciones: 1. Comparar una medida de ajuste del modelo obtenida bajo cada distribución candidata, 2. Evaluar las estimaciones de los parámetros del modelo *ZIP* obtenidas bajo cada distribución candidata; y 3. Contaminar los datos de cultivo de manzanas con una observación atípica y observar bajo cuál distribución candidata se realiza un mejor ajuste de la misma.

En la tabla 2 se presenta la medida de ajuste obtenida para el modelo *ZIP* con cada una de las distribuciones candidatas como a priori para sus parámetros de escala, la medida de ajuste presentada es el DIC (Deviance information criterion), la cual es un criterio de información que evalúa el ajuste de un modelo, que a su vez penaliza la complejidad del mismo, entre múltiples modelos se prefiere aquel de menor DIC, además se dirá que existe una diferencia significativa entre el ajuste ofrecido por dos modelos si la diferencia entre los DIC calculados para cada uno es mayor o igual a 5. Se observa que en general los valores de la medida de ajuste obtenida con cada candidata son muy cercanos entre sí, la diferencia entre los DIC es menor a 5, por lo cual se concluye que no existe una diferencia marcada.

En la tabla 3 se presentan las estimaciones a posteriori de los principales parámetros del modelo *ZIP*, obtenidas con cada una de las distribuciones candidatas,

Tabla 2: *Medida de ajuste - Distribuciones candidatas. Fuente: elaboración propia.*

Distribución	DIC
Gamma inversa	1873
SBeta2	1870
Half Cauchy	1871

mientras que en la tabla 4 se presentan las estimaciones de los parámetros de escala. La estimación para un determinado parámetro, bajo una candidata específica, se obtuvo a partir de la mediana de la cadena a posteriori de dicho parámetro. Se observa que las estimaciones obtenidas para cada uno de los parámetros del modelo *ZIP* son muy similares entre las diferentes distribuciones candidatas. Esto puede explicarse en cuanto a la gran cantidad de información muestral disponible (270 datos): es decir, que las distribuciones a posteriori están más influenciadas por la información muestral.

Tabla 3: *Estimación parámetros principales - Modelo ZIP. Fuente: elaboración propia.*

Distribución	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\gamma_0}$	$\widehat{\gamma_1}$	$\widehat{\gamma_2}$
Gamma inversa	1.98	0.1055	0.05003	-2.434	0.152	-0.053
SBeta2	1.98	0.1057	0.05015	-2.438	0.153	-0.056
Half Cauchy	1.98	0.1058	0.04982	-2.436	0.144	-0.050

Tabla 4: *Estimación parámetros de escala - Modelo ZIP. Fuente: elaboración propia.*

Distribución	$\widehat{\sigma_1^2}$	$\widehat{\sigma_2^2}$	$\widehat{\sigma_3^2}$	$\widehat{\sigma_4^2}$
Gamma inversa	0.0248	0.0332	0.0671	0.0743
SBeta2	0.0408	0.0612	0.0842	0.0941
Half Cauchy	0.0301	0.0505	0.0784	0.0875

En la tabla 5 se presentan intervalos de credibilidad al 95 % para los principales parámetros del modelo *ZIP*, obtenidos con cada una de las distribuciones candidatas, mientras que en la tabla 6 se presentan los intervalos de credibilidad para los parámetros de escala. Se observa que los intervalos de credibilidad para cada uno de los parámetros son muy similares entre si con las diferentes candidatas, en cuanto a la longitud de los mismos y el rango de valores en los que fluctúan. Esto nuevamente puede verse explicado por la gran cantidad de información muestral disponible, así como a la predominancia que esta debe tener sobre las distribuciones a posteriori.

La tabla 7 presenta los resultados del ajuste de una observación atípica con cada una de las distribuciones candidatas. Se observa que las distribuciones candidatas que ofrecen un mejor ajuste de la observación atípica son la SBeta2 y la half-Cauchy, con valores relativamente cercanos. La distribución Gamma inversa ofrece un peor ajuste de la observación atípica.

Tabla 5: *Intervalos de Credibilidad al* 95 % *- Parámetros principales. Fuente: elaboración propia.*

Distribución	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$
Gamma inversa	(1.907,1.999)	(0.098,0.113)	(0.042,0.058)
SBeta2	(1.902,1.999)	(0.097,0.113)	(0.043,0.057)
Half Cauchy	(1.895,1.999)	(0.098,0.114)	(0.041,0.059)
Distribución	$\widehat{\gamma_0}$	$\widehat{\gamma_1}$	$\widehat{\gamma_2}$
Gamma inversa	(-2.497,-2.184)	(0.117,0.184)	(-0.094,-0.005)
SBeta2	(-2.499,-2.109)	(0.121,0.184)	(-0.089,-0.014)
Half Cauchy	(-2.498,-2.113)	(0.113,0.182)	(-0.099,-0.007)

Tabla 6: *Intervalos de credibilidad al* 95 % *- Parámetros de escala. Fuente: elaboración propia.*

Distribución	$\widehat{\sigma_1^2}$	$\widehat{\sigma_2^2}$	$\widehat{\sigma_3^2}$	$\widehat{\sigma_4^2}$
Gamma inversa	(0.0045,0.5781)	(0.0062,0.9263)	(0.0086,1.3462)	(0.0092,1.5782)
SBeta2	(0.0072,0.9801)	(0.0122,1.6948)	(0.0097,1.6543)	(0.0095,1.8123)
Half Cauchy	(0.0027,0.8580)	(0.0044,1.4931)	(0.0068,0.9254)	(0.0073,1.3284)

5. Conclusiones

Dentro de las condiciones en las que se enmarcó el estudio de simulación presentado en este artículo, para la distribución Gamma inversa se evidencia de manera más fuerte el problema de encogimiento en las estimaciones a posteriori de los parámetros. Dicha situación mejora considerablemente con las distribuciones half-Cauchy y SBeta2, lo que las hace más recomendables como a priori para los parámetros de escala del modelo *ZIP*. Además, la distribución Gamma inversa presenta más dificultad a la hora de ajustar observaciones atípicas. Para las distribuciones half-Cauchy y SBeta2 mejora considerablemente la capacidad del modelo de ajustar observaciones atípicas,se obtuvierón resultados similares con las dos distribuciones, aunque la SBeta2 bajo cualquier escenario ofrece una leve mejora. Con esto se obtiene otro atributo que hace mucho más recomendable las distribuciones half-Cauchy y SBeta2 como a priori para los parámetros de escala del modelo *ZIP*, por encima de la distribución Gamma inversa.

De los resultados del caso práctico se puede concluir que bajo las condiciones delcaso, las distribuciones candidatas consideradas como a priori para los parámetros

Tabla 7: *Ajuste de una observación atípica - Distribuciones candidatas. Fuente: elaboración propia.*

Distribución	Ajuste	Real	Diferencia absoluta
Gamma inversa	18.66	34	15.34
SBeta2	26.61	34	7.39
Half Cauchy	25.66	34	8.34

de escala del modelo *ZIP* presentan entre ellas un ajuste del modelo relativamente similar. Además, que las estimaciones obtenidas con cada candidata son cercanas. Sin embargo, las distribuciones SBeta2 y half-Cauchy ofrecen un mejor ajuste de una observación atípica que el que ofrece la distribución Gamma inversa.

Recibido: 14 de mayo de 2016
Aceptado: 27 de septiembre de 2016

Referencias

Barrera, C. & Correa, J. (2008), 'Distribución predictiva bayesiana para modelos de pruebas de vida vía MCMC', *Revista Colombiana de Estadística* **31**(2), 145–155.

Berger, J. (2006), 'The case for objective Bayesian analysis', *Bayesian Analysis* **1**(3), 385–402.

Fruhwirth-Schnatter, S. & Wagner, H. (2010), 'Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data', *Bayesian Statistics* **9**, 165.

Fúquene, J., Pérez, M. & Pericchi, L. (2014), 'An alternative to the Inverted Gamma for the variances to modelling outliers and structural breaks in dynamic models', *Brazilian Journal of Probability and Statistics* **28**(2), 288–299.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models', *Bayesian Analysis* **1**(3), 515–533.

Ghosh, S., Mukhopadhyay, P. & Lu, J. (2006), 'Bayesian analysis of zero-inflated regression models', *Journal of Statistical Planning and Inference* **136**, 1360–1375.

Heibron, D. (1994), 'Zero-altered and other regression models for count data with added zeros', *Biometrical Journal* **36**, 531–547.

Kwiatkowski, D., Phillips, P. & Schmidt, P. (1992), 'Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root', *Journal of Econometrics* **54**, 159–178.

Lambert, D. (1992), 'Zero-Inflated Poisson regression with an application to defects in manufacturing', *Technometrics* **34**, 1–14.

Marin, J., Jones, O. & Hadlow, W. (1993), 'Micropropagation of columnar apple trees', *Journal of Horticultural Science* **68**(2), 289–297.

O'Hagan, A. & Pericchi, L. (2012), 'Bayesian heavy-tailed models and conflict resolution: A review', *Brazilian Journal of Probability and Statistics* **26**(4), 372–401.

Pérez, M., Pericchi, L. & Ramírez, I. (2016), 'The Scaled Beta2 distribution as a robust prior for scales', *Artículo sometido para Publicación* .

Pericchi, L. (2010), 'Discussion of Polson, N., and Scott, J.', *Bayesian Statistics* **9**, 531.

Polson, N. & Scott, J. (2012), 'On the half-Cauchy prior for a global scale parameter', *Bayesian Analysis* **7**(4), 887–902.

Rodrigues, J. (2006), 'Full Bayesian Significance Test for Zero-Inflated Distributions', *Communications in Statistics - Theory and Methods* **35**(2), 299–307.