UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2014, Vol. 7, No. 2, pp. 115–129

Análisis de distribuciones a priori de 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 (conocido como modelo ZIP por sus siglas en inglés). 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; y la beta 2 escalada (SBeta2) la cual es una distribución de colas pesadas que tiene un mejor comportamiento 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 de los parámetros a posteriori del modelo; además se evalúa ante la presencia de observaciones atípicas cómo es el ajuste que el modelo realiza de estas, con cada una de las distribuciones a priori candidatas para los parámetros de escala. El análisis se centra en estas dos características (encogimiento de los parámetros a posteriori 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]Escuela de Estadística - Universidad Nacional de Colombia, sede Medellín
[b]Escuela de Estadística - Universidad Nacional de Colombia, sede Medellín

Abstract

In this paper, It is propose the evaluation of a set of prior distributions for the scales parameters of the Zero-Inflated Poisson Regression model (ZIP). Traditionally the inverse-gamma distribution is used as prior for scales parameters. Some studies have shown that when the values of the hyperparameters of this distribution are very small, subsequent inferences are not adequate. Our focus is on evaluating three priors for model's scales parameters: inverted gamma; the Half Cauchy that has been used to the situation in question and that has proven to work properly; and scaled beta 2 (SBeta2) which is a heavy-tailed distribution that has a better performance at the origin and at the right tailed.

A simulation study is developed, with which we intend to analyze the effect of the prior distribution assigned to the scales parameters on the shrinkage of the posterior model's parameters; also is evaluated with the presence of outliers how the model performs adjustment of these, for each of the candidates prior distributions for the parameters of scale. The analysis focuses on these two characteristics (shrinkage of the posterior parameters and adjustment of outliers) because these are the main criticisms different authors suggest to the use of inverse-gamma distribution as a priori for parameters of scale. Finally is presented an application 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: 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, por ejemplo Berger (2006) 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, lo cual puede conllevar a su

vez a resultados incoherentes y la incapacidad de predecir o ajustar observaciones atípicas; por su parte Gelman (2006) argumenta que la gamma-inversa(ϵ, ϵ) cuando se usa como a priori para la varianza, buscando que sea no informativa se hace $\epsilon \to 0$, lo cual en realidad produce un encogimiento en los parámetros a posteriori del modelo, y si por la naturaleza de los datos es posible valores pequeños de la varianza, la a priori se convierte en informativa, además el autor ilustra por medio de un ejemplo con datos reales el problema de concentración alrededor del cero.

La segunda alternativa a 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 tercer alternativa es la distribución Beta2 escalada (SBeta2) propuesta para este uso por Pericchi (2010), para la cual se sabe posee unas propiedades teóricas convenientes cuando se usa como a priori para parámetros de escala.

Para 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 cual para cada distribución a priori candidata se analizó las condiciones de encogimiento de los parámetros a posteriori 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 (por ejemplo (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 así: 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, por ejemplo como la distribución Poisson. Sin embargo, existen casos en 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 habla 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 comúnmente utilizado para trabajar datos de conteo con exceso de ceros, el cual fue propuesto por Lambert (1992). Bajo este modelo se tiene 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:

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

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 $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ 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. Además, 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 y para valores suficientemente grandes de su hiperparámetro es una a priori débilmente informativa, además 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 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 3 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 36 escenarios, conformados 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: $\sigma_1^2 = \sigma_2^2 = 0.1, 3, 10, 35$; 3 tamaños muestrales: $n = 5, 15, 30$. 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: Primero, para una determinada distribución candidata, se construyó el código del modelo ZIP, ajustando la distribución candidata como a priori para sus parámetros de escala. Segundo, 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. Tercero, 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), buscando 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 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 adecuado comportamiento 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 los parámetros a posteriori.

Se presentan resultados para cuatro condiciones principales: fijando los parámetros de escala en $\sigma_1^2 = \sigma_2^2 = 0.1, 3, 10, 35$.

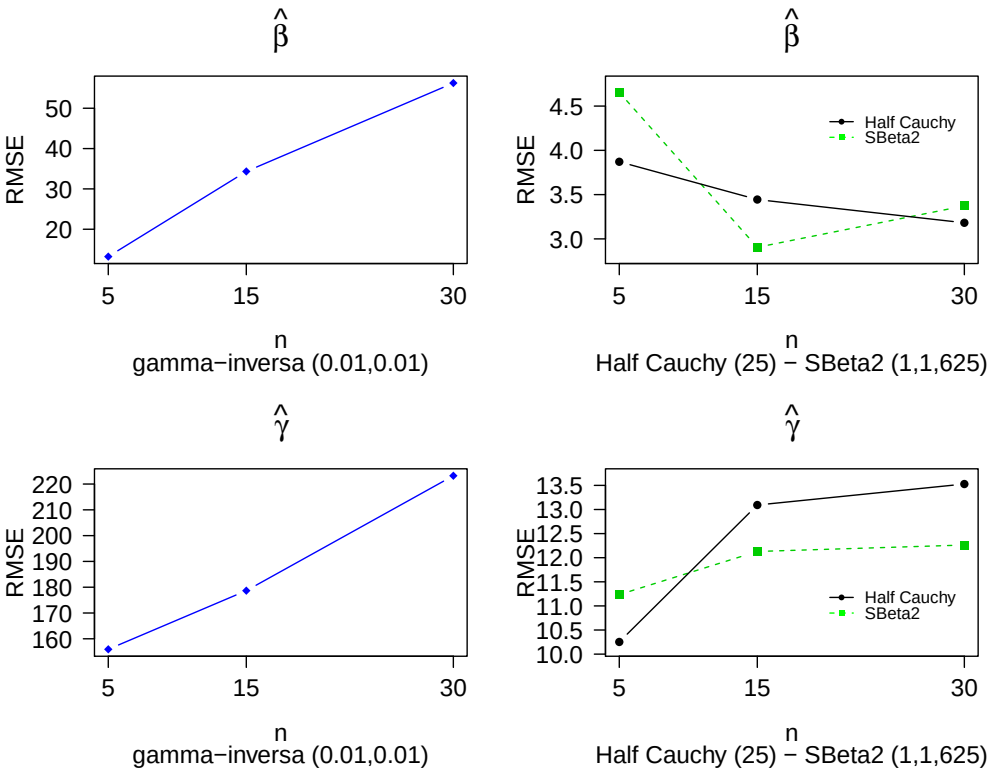

Figura 1: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 0.1$

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 se hace necesario realizar un chequeo de convergencia sobre las cadenas a posteriori obtenidas en este estudio de simulación. En

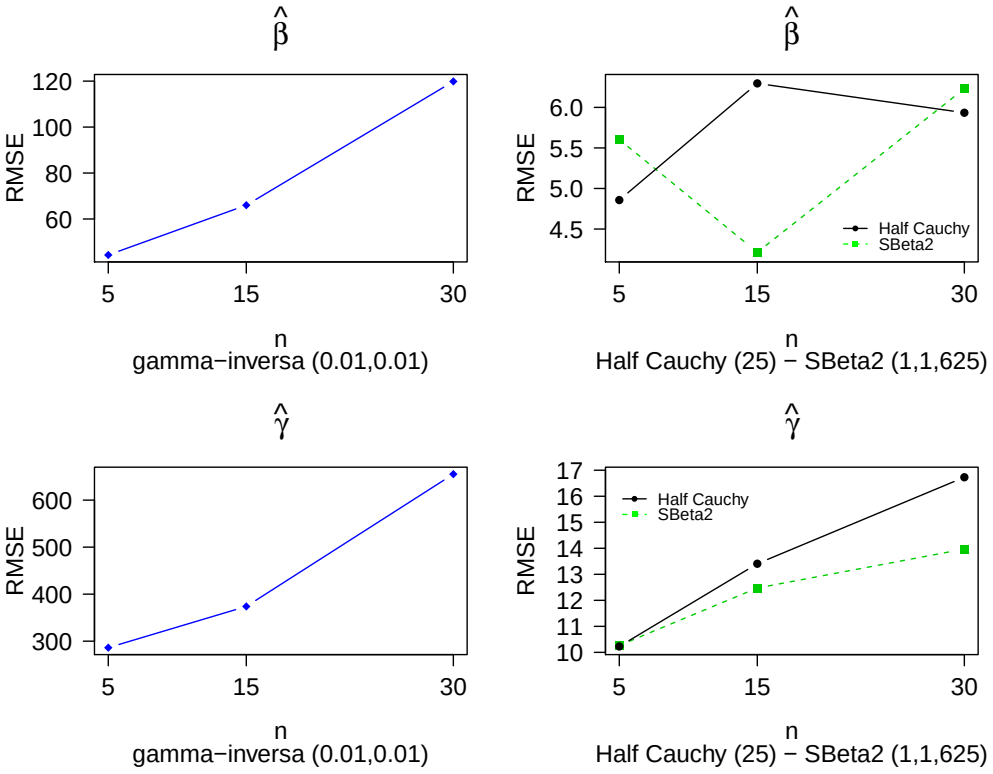

Figura 2: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 3$

el presente trabajo el chequeo de convergencia se realiza de forma similar al procedimiento planteado para dicho fin por Barrera & Correa (2008). Así, para una determinada cadena, el chequeo consiste en evaluar la autocorrelación existente entre los valores generados del parámetro en distintos rezagos; se realiza un gráfico de promedios móviles y por último se realiza 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:

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

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).

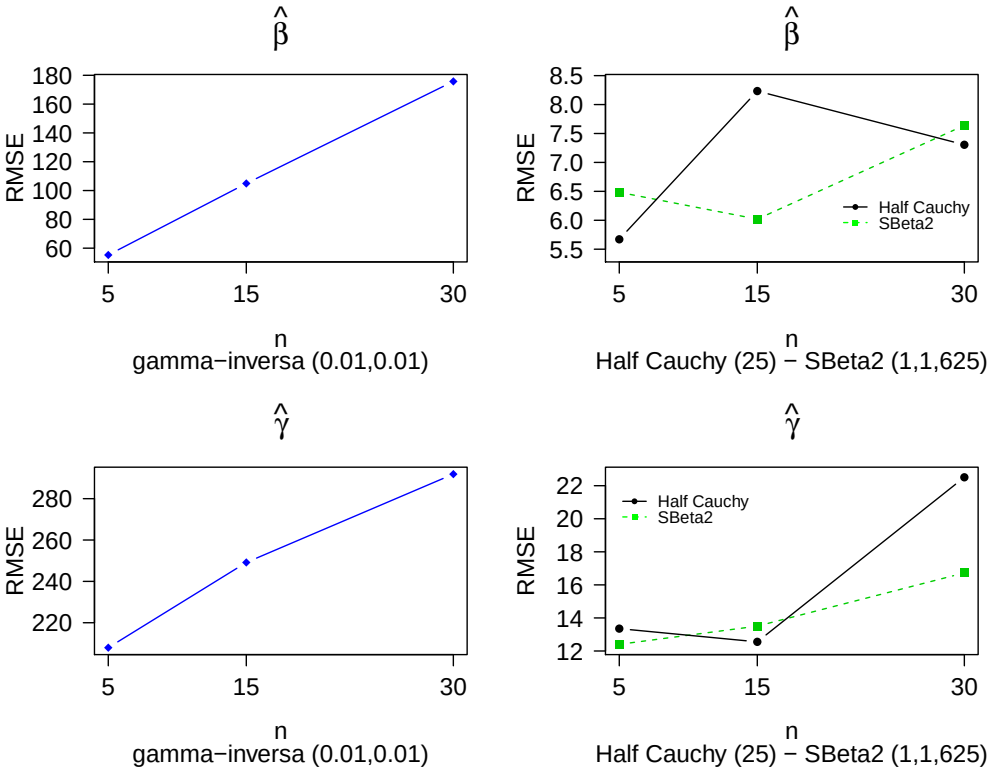

Figura 3: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 10$

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

	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 con-

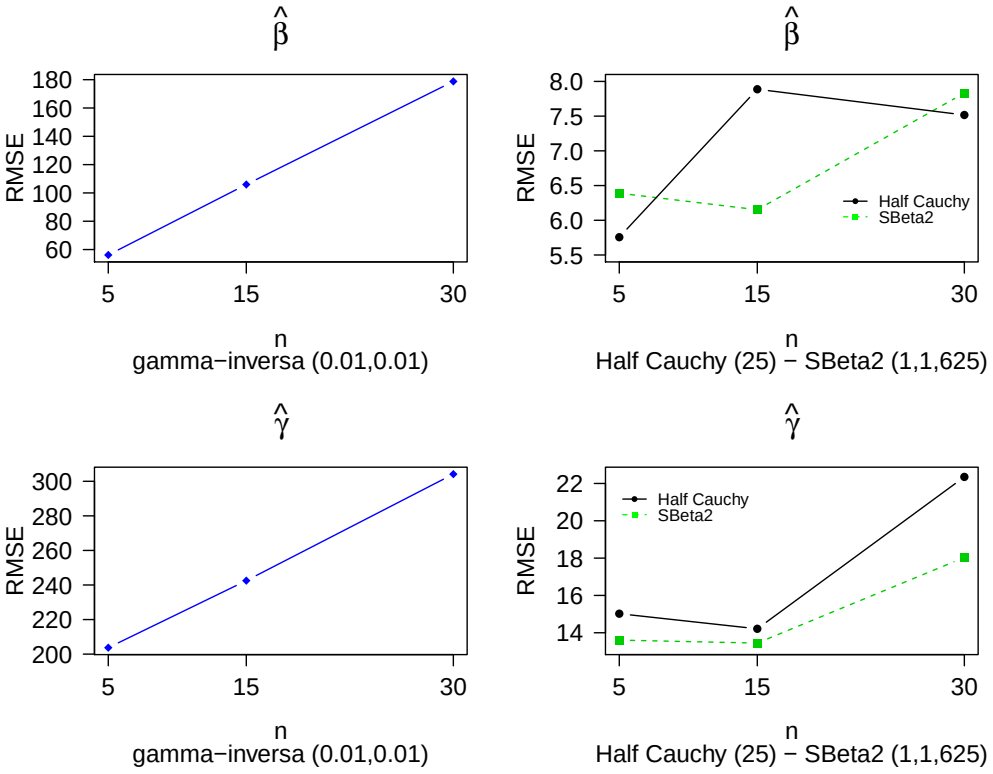

Figura 4: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 35$

cluye 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.

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 es usada como a priori para los parámetros de escala, en cuanto al inadecuado ajuste de observaciones atípicas que ocurre, esto ya que con

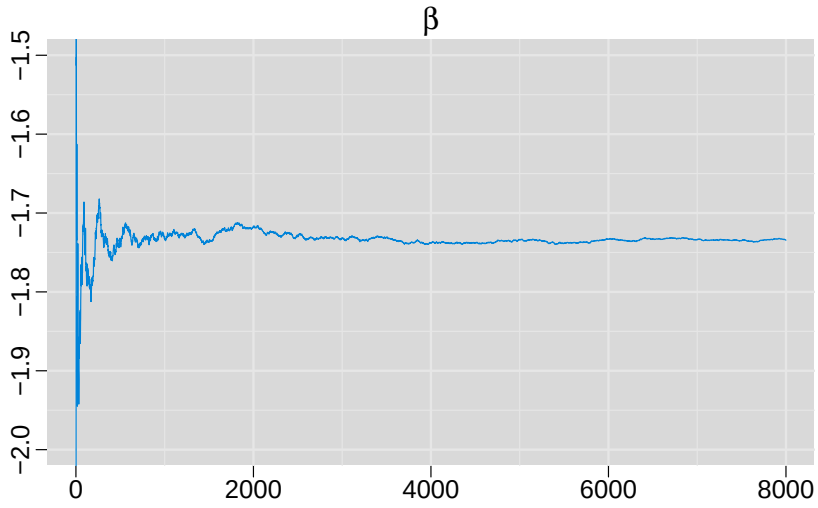

Figura 5: Promedios móviles - Cadena del análisis de encogimiento

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 determinada 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.

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

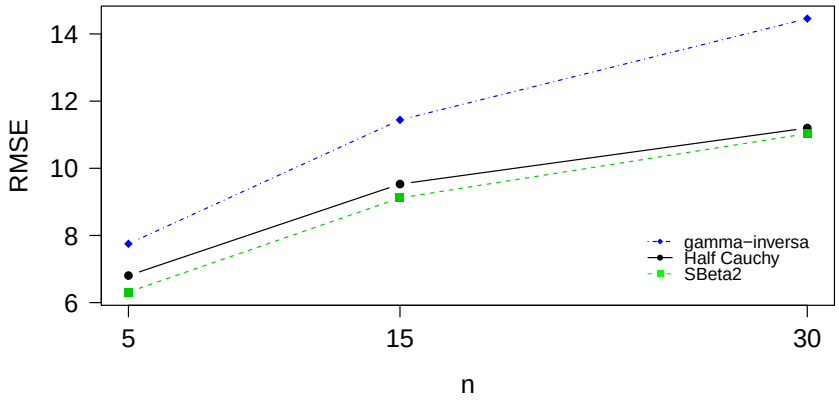

Figura 6: RMSE VS Tamaño muestral - Ajuste de una observación atípica

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:

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

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: Comparar una medida de ajuste del modelo obtenida bajo cada distribución candidata; Evaluar las estimaciones de los parámetros del modelo ZIP obtenidas bajo cada distribución candidata; y 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), el cual es un criterio de información que evalúa el ajuste de un modelo, penalizando a su vez 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 si, la diferencia entre los DIC es menor a 5, por lo cual se concluye que no existe una diferencia marcada.

Tabla 2: Medida de ajuste - Distribuciones candidatas

Distribución	DIC
gamma-inversa	1873
SBeta2	1870
Half Cauchy	1871

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

La tabla 4 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 3: Estimación parámetros - Modelo ZIP

Distribución	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$
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: Ajuste de una observación atípica - Distribuciones candidatas

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

5. Conclusiones

Dentro de las condiciones en 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 de los parámetros a posteriori. 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, obteniéndose resultados similares con las dos distribuciones, aunque la SBeta2 bajo cualquier escenario ofrece una leve mejora. Teniéndose así otro atributo que hace mucho más recomendables 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 bajos las condiciones de este, las distribuciones candidatas consideradas como a priori para los parámetros 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:
Aceptado:

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.