Análisis de distribuciones a priori de los parámetros de escala del modelo de regresión Poisson inflado con ceros

Juan Daniel Molina Muñoz

Universidad Nacional de Colombia
Sede Medellín
Facultad de Ciencias
Escuela de Estadística
Medellín, Colombia
2016

Análisis de distribuciones a priori de los parámetros de escala del modelo de regresión Poisson inflado con ceros

Juan Daniel Molina Muñoz

Trabajo presentado como requisito parcial para optar al título de:
Magister en Ciencias - Estadística

Director:
Isabel Cristina Ramírez Guevara, Ph.D.(c) en Estadística

Universidad Nacional de Colombia
Sede Medellín
Facultad de Ciencias
Escuela de Estadística
Medellín, Colombia
2016

A Dios quien es la fuente de mi fortaleza, a mi familia cuyo amor es un faro en mi vida.

Agradecimientos

A mi directora de tesis la profesora Isabel Cristina Ramírez Guevara por su orientación y acompañamiento durante el desarrollo de este trabajo. A mis jurados de tesis, los profesores Juan Carlos Correa Morales y Freddy Hernández Barajas, cuyos comentarios y recomendaciones fueron de gran ayuda para el enriquecimiento de esta tesis. A los profesores de la escuela de Estadística de la universidad Nacional de Colombia sede Medellín y compañeros de la maestría en Estadística que aportaron a mi formación.

A mis padres por el apoyo que siempre me han dado.

Resumen

En el presente trabajo 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 de cultivo de manzanas.

Palabras clave: Inferencia Bayesiana, Modelo ZIP, Parámetros de escala, Distribución SBeta2, Distribución Half Cauchy, Distribución gamma-inversa.

Abstract

In this thesis, 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 like 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 adjust-

ment 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 like a priori for parameters of scale. Finally is presented an application with real data of growing apples.

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

Contenido

Lista de Tablas

Lista de Figuras

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 de regresión Possion inflado con ceros (conocido como modelo ZIP por sus siglas en inglés) 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 unas 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 varianza 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 de parámetros de escala, lo que hace interesante evaluar dicha alternativa.

Para evaluar el impacto de la distribución a priori asignada a los parámetros de escala en el modelo ZIP se realizó un estudio 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. Además se presenta una aplicación con datos reales de cultivo de manzanas, los cuales fueron obtenidos por Marin et al. (1993).

Los siguientes capítulos del presente documento están organizados así: en el capítulo 2 se presenta la definición del modelo de regresión Poisson inflado con ceros (ZIP), la caracterización de la aplicación del modelo ZIP desde un enfoque Bayesiano, algunas propuestas que se han planteado para la distribución a priori de los parámetros de escala, se enuncian las características de la distribución SBeta2 y las ventajas que esta distribución presenta al usarse como a priori de parámetros de escala. En el capítulo 3 se desarrolla el estudio de simulación, se realiza una definición de las características generales y condiciones del mismo, se presentan y analizan los resultados referente al fenómeno de encogimiento de los parámetros a posteriori del modelo y referente al ajuste de observaciones atípicas. En el capítulo 4 se desarrolla un caso práctico con datos reales de cultivo de manzanas. Finalmente en el capítulo 5 se presentan las principales conclusiones obtenidas de este trabajo y se mencionan algunos posibles trabajos futuros de investigación.

2. Marco teórico

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.

Los modelos que trabajan con un exceso de ceros han sido utilizados en una gran cantidad de fenómenos de conteo de la vida real. Por ejemplo, Lambert (1992) define el modelo de regresión Poisson inflado con ceros (ZIP) y presenta una aplicación del mismo con datos de defectos de fabricación; Heibron (1994) plantea un modelo de regresión cero-alterado para trabajar datos de conteo con una cantidad excesiva de ceros y presenta una aplicación con datos relacionados con un fenómeno de comportamiento sexual. La diferencia entre los modelos inflados con ceros y los modelos cero-alterado está dada en que en los modelos inflados, los ceros son generados por dos fuentes: ceros extras y ceros provenientes de una determinada distribución, en cambio en los modelos cero-alterado todos los ceros se asumen ceros extras; Cheung (2002) presenta una aplicación de modelos de regresión inflados con ceros en un contexto médico, con datos relacionados con el crecimiento y desarrollo de niños; Agarwal et al. (2002) presentan una aplicación del modelo ZIP en el contexto espacial; Famoye & Singh (2006) presentan una aplicación del modelo ZIP generalizado con datos de violencia doméstica; Karlis & Ntzoufras (2003) proponen una versión bivariada del modelo ZIP para pronosticar campeonatos deportivos.

2.1. Modelo de regresión Poisson inflado con ceros (ZIP)

Este 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 $(n, k+1)$.

2.1.1. Aplicación del modelo ZIP desde un enfoque Bayesiano

Ghosh et al. (2006) presentan un procedimiento para modelar datos provenientes de un fenómeno de conteo con la característica de una presencia excesiva de ceros. Se plantea una familia general de distribuciones para dicha situación conocida como "modelos de series de potencias infladas con cero", a la cual pertenece el modelo ZIP y para el cual se presenta una aplicación. Plantean además un enfoque Bayesiano para la estimación de los parámetros de estos modelos y se hace una comparación de dicho enfoque con el realizado vía máxima verosimilitud o clásico. Los autores concluyen que el enfoque Bayesiano arroja mejores resultados cuando se tienen muestras pequeñas, o cuando la proporción de ceros en los datos es alta (cercana al 100 %). La aplicación de la metodología propuesta se ilustra a partir de muestreo vía simulación por medio del software WinBUGS® usando datos provenientes de un contexto de manufactura.

Otros autores que abordan modelos de regresión inflados con ceros desde un enfoque Bayesiano son Angers & Biswas (2003), los cuales plantean un conjunto de posibles distribuciones a priori para los parámetros del modelo ZIP generalizado y presentan una aplicación en datos reales relacionados con movimientos fetales de corderos; Rodrigues (2003) plantea

el uso del algoritmo *data augmentation* en distribuciones infladas con cero y presenta una aplicación en el caso del modelo ZIP a partir de datos relacionados con el cultivo de manzanas; Xie et al. (2014) presentan consideraciones generales para abordar el modelo ZIP generalizado desde un enfoque Bayesiano y muestran una aplicación con datos provenientes de un experimento psicológico.

2.1.2. Caracterización de la metodología Bayesiana

Un paso fundamental en el desarrollo de una metodología Bayesiana es la determinación de las distribuciones a priori de los parámetros del modelo, en el caso del presente trabajo el interés se centra en determinar las distribuciones a priori para los parámetros de escala en el modelo ZIP. Para esto existen dos opciones básicas: el análisis Bayesiano subjetivo o el objetivo. El análisis subjetivo utiliza el proceso de elicitación en la construcción de las distribuciones a priori. Por su parte el análisis objetivo trata de utilizar como a priori distribuciones objetivas, es decir, distribuciones que tengan un mínimo impacto informativo sobre el análisis y que se ajusten al contexto de los datos. Berger (2006) hace una descripción general de las técnicas del análisis Bayesiano objetivo, el por qué de su amplio uso en la actualidad, algunas críticas realizadas a las mismas y sus argumentos para considerar que estas técnicas son realmente confiables. En este trabajo se pretende hacer un análisis objetivo.

2.2. Propuestas de distribuciones a priori para parámetros de escala

A continuación se presentan algunas de las propuestas que existen para las distribuciones a priori de los parámetros de escala en modelos jerárquicos, ya que de manera general, dentro del enfoque Bayesiano comúnmente se consideran los parámetros de escala como parte de una estructura jerárquica. Daniels (1999) plantea que comúnmente como a priori de la varianza en modelos jerárquicos se usa la distribución gamma-inversa, por ser la conjugada de la distribución normal. Sin embargo, en muchas aplicaciones muy poca o ninguna información previa sobre los componentes de varianza está disponible (en aquellas situaciones que se considera un modelo de tipo ANOVA), en estos casos, una distribución a priori "vaga" o "no informativa" es deseada para reflejar dicho desconocimiento. Una opción es usar la distribución gamma-inversa asumiendo que la varianza de la misma es muy grande. Sin embargo, al hacer ésto se produce una distribución a priori "impropia" que conduce a su vez a una distribución a posteriori impropia, en especial cuando se trabajan con muestras pequeñas. Así Daniels (1999) propone una distribución a priori "vaga", denominada "uniform shrinkage", además presenta sus propiedades y muestra que si es usada en modelos jerárquicos conduce a una distribución a posteriori propia.

Por su parte Gustafson et al. (2006) proponen el uso de distribuciones a priori "conservadoras" para los componentes de varianza dentro de los modelos jerárquicos, las cuales deliberadamente le dan más peso a los valores pequeños. Esta opción puede ser adecuada para los investigadores que son escépticos acerca de la presencia de variabilidad en los parámetros de la segunda etapa (efectos aleatorios) y desean evitar una estructura en el modelo mayor a la realmente presente. Los autores plantean que las distribuciones a priori sugeridas se adaptan fácilmente a diversos ajustes del modelamiento jerárquico, como curvas suaves de ajuste, modelamiento de la variación espacial y la combinación de datos de múltiples contextos.

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. Se ilustra también el uso de la familia de distribuciones half-t para la modelación jerárquica de múltiples parámetros de varianza provenientes de un análisis de varianza, y específicamente 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 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, lo cual es deseado en el desarrollo de un análisis objetivo. Como estas a priori representan una mezcla de escalas, son adecuadas para detectar cambios repentinos en un modelo. Los autores proponen además utilizar esta nueva familia de distribuciones a priori dentro del esquema del muestreador de Gibbs, para modelar observaciones atípicas y cambios estructurales en los modelos lineales dinámicos bayesianos. Por medio de un ejemplo los autores muestran que la propuesta funciona mejor que el caso en que se asigna como a priori para la varianza la distribución gamma-inversa, pues de esta forma es más difícil detectar cambios estructurales.

2.2.1. Distribución SBeta2 como a priori para los parámetros de escala

La evaluación de la distribución SBeta2 como a priori para los parámetros de escala del modelo ZIP, representa uno de los principales aportes metodológicos de esta tesis, pues para las otras dos candidatas ya existen estudios en que se evalúan como a priori para parámetros de escala. Por tanto, en la presente sección se detallan las características de la distribución SBeta2. Se parte de la definición de la distribución Beta2, también conocida como distribución "Beta prime" o "Beta-inversa", la cual está determinada por la siguiente función de densidad de probabilidad:

$$Beta2(\psi|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\frac{\psi^{p-1}}{(\psi+1)^{p+q}}; \quad \psi>0, p>0, q>0,$$

donde $\psi = \frac{\varpi}{1-\varpi}$, conocido como el *odd ratio* y $\varpi \sim Beta(\varpi|p,q)$.

La SBeta2 es una versión escalada de la Beta2, de esta forma la distribución SBeta2 para el parámetro ψ se define a partir de la siguiente función de densidad de probabilidad:

$$SBeta2(\psi|p,q,b) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)b}\frac{(\frac{\psi}{b})^{p-1}}{(1+\frac{\psi}{b})^{p+q}}; \quad \psi>0, p>0, q>0, b>0. \tag{2-1}$$

Además, la distribución SBeta2 presenta las siguientes formulas de esperanza y varianza:

$$\begin{aligned} E[\psi] &= \frac{p}{q-1}b \quad ; \quad \text{cuando} \quad q>1, \\ Var[\psi] &= \frac{p(p+q-1)}{(q-1)^2(q-2)}b^2 \quad ; \quad \text{cuando} \quad q>2. \end{aligned}$$

Pericchi (2010) propone el uso de la distribución SBeta2 como un remplazo de la gamma-inversa como a priori para los parámetros de escala en modelos jerárquicos. En Pericchi & Pérez (2009) se comprueba la robustez de la distribución SBeta2 por medio de la teoría de *Regularly Varying functions*. Además, los autores muestran que la distribución SBeta2 puede definirse como una mezcla de gammas para el cuadrado de la escala, de la siguiente forma:

$$\begin{aligned} \psi^2 &\sim \text{gamma}(p, b/\rho), \\ \rho &\sim \text{gamma}(q, 1), \end{aligned}$$

donde gamma(α,β) denota la distribución gamma con función de densidad de probabilidad dada por:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp(-x/\beta); \quad x>0, \alpha>0, \beta>0.$$

Fúquene et al. (2014) modelan el cuadrado de la escala de la t-student con una a priori SBeta2 y muestran que la marginal de localización puede escribirse de forma explícita. De esta manera se obtiene una a priori de colas pesadas que en general puede usarse para los parámetros de escala en el análisis Bayesiano. Esta metodología se adapta naturalmente al procedimiento de muestreo simple de Gibbs, sin añadir alguna complicación respecto al análisis de la a priori gamma-inversa, pero si con una mejora en el rendimiento.

De una manera más general Pérez et al. (2015) 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, capaz de modelar diferentes tipos de comportamientos en el origen y la cola. La SBeta2 se rige por 3 parámetros (como se observa en la **ecuación** 2-1) que son factibles de elicitar, p rige el comportamiento en el origen, q el de la cola derecha y b la escala de la distribución. Si se define una a priori condicional normal o Cauchy para la localización y la SBeta2 para la escala, es posible obtener una a priori cerrada para la marginal de localización. En especial, en el caso en que la a priori condicional para la localización es Cauchy, para determinados valores de los hiperparámetros se llega a una marginal de localización conocida como *explicit Horseshoe*, que posee un polo en el origen y colas pesadas, la cual es de gran utilidad en modelos Bayesianos jerárquicos completos. Finalmente, la SBeta2 es una distribución robusta, donde el espesor de su cola es equivalente al de la t-student (Pérez et al. 2015).

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, pues de lo contrario dicha proporción sería igual siempre a 0.5 (Hardin & Hilbe 2007). De esta forma, el modelo se resume en la siguiente expresión:

$$\begin{aligned} &Y_i \sim \text{ZIP}(p_i, \lambda_i), \\ &\log(\lambda_i) = \beta X, \\ &\text{logit}(p_i) = \gamma_0 + \gamma X. \end{aligned} \tag{3-1}$$

Se asume que $\beta \sim N(0, \sigma_1^2)$, $\gamma \sim N(0, \sigma_2^2)$ y $\gamma_0 \sim U(-2.5, 2.5)$ pues de esta forma se garantiza que la proporción extra de ceros tome valores en 0 y 1 (Hardin & Hilbe 2007); durante el estudio de simulación se asumió $X \sim U(0, 1)$ además σ_1^2 y σ_2^2 fueron valores fijos dentro de la simulación.

Dentro del presente estudio de simulación 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 presente 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. La sintaxis del código del modelo se construyó para que este sea ejecutado en el software bayesiano OpenBUGS®. Este procedimiento se realizó con cada una de las distribuciones candidatas. Los códigos del ajuste del modelo ZIP con cada una de las distribuciones candidatas se incluye en los anexos.

Segundo, para un determinado escenario de simulación, se generaron los datos de la variable respuesta que distribuye ZIP y de la covariable, para este procedimiento se utilizó el

software R. Como ya se había mencionado para la covariable se asume $X \sim U(0,1)$. Para generar los valores de la variable respuesta se parte de la estructura del modelo presentada en 3-1, así, a partir de los valores establecidos para σ_1^2 y σ_2^2 (dado el escenario a trabajar), se generan β y γ, sabiendo que $\beta \sim N(0, \sigma_1^2)$ y $\gamma \sim N(0, \sigma_2^2)$, adicionalmente se genera γ_0 sabiendo que $\gamma_0 \sim U(-2.5, 2.5)$, después con los valores de β, γ y γ_0 se definen los valores de λ_i y p_i, pues de 3-1 se tiene que $\lambda_i = \exp(\beta X)$ y para hallar p_i se usa la función inversa del logit sobre $\gamma_0 + \gamma X$ (para esto puede usarse la función *inv.logit* de la librería *boot* del software R), finalmente con los valores de λ_i y p_i se generan los valores de la variable respuesta, sabiendo que $Y_i \sim \text{ZIP}(p_i, \lambda_i)$, para esto puede usarse la función *rZIP* de la librería *gamlss.dist* del software R. El código de la generación de los valores de la variable respuesta y la covariable se incluye en los anexos.

Tercero, para un determinado escenario, se simularon las cadenas a posteriori de los parámetros del modelo ZIP, esto se hace por medio del método MCMC (Markov Chain Monte Carlo), el método MCMC es un algoritmo de muestreo que de forma aleatoria acepta valores para la distribución a posteriori de los parámetros de un modelo, el procedimiento de generación de las cadenas a posteriori de los parámetros del modelo ZIP se realizó por medio del software OpenBUGS®, el cual usa como insumos el modelo ZIP ajustado con una determinada distribución candidata como a priori para sus parámetros de escala y los datos generados de la variable respuesta y de la covariable. La simulación de las cadenas a posteriori se realizaron de tal manera que para cada uno de los parámetros del modelo se generaran 10000 valores del mismo, con un quemado inicial de 2000 valores, es decir, que finalmente se dispone de 8000 valores a posteriori generados por parámetro. El código de la generación de las cadenas a posteriori para los parámetros del modelo ZIP se incluye en los anexos.

Finalmente, es importante mencionar que dentro del estudio de simulación 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. (2015) muestran que la Half Cauchy(25^2) que puede usarse como a priori para la varianza es equivalente a la SBeta2$(0.5, 0.5, 25^2)$. Finalmente, Pérez et al. (2015) muestran que la SBeta2$(1, 1, 25^2)$ presenta un adecuado comportamiento cuando es usada como a priori para la varianza.

3.1. Análisis de encogimiento de los parámetros a posteriori del modelo

Como ya se ha mencionado, algunos autores plantean que al usar la distribución gamma-inversa como a priori para los parámetros de escala se incurre en problemas como el encogimiento de los parámetros a posteriori del modelo (Gelman 2006). Dada esta situación, 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). Para esto, 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 del parámetro a posteriori.

Los resultados del análisis de encogimiento se presentarán para cuatro condiciones principales: fijando los parámetros de escala en $\sigma_1^2 = \sigma_2^2 = 0.1$; $\sigma_1^2 = \sigma_2^2 = 3$; $\sigma_1^2 = \sigma_2^2 = 10$ y $\sigma_1^2 = \sigma_2^2 = 35$.

En el caso de $\sigma_1^2 = \sigma_2^2 = 0.1$ (**Figura 3-1**), tanto para $\hat{\beta}$ como para $\hat{\gamma}$ se observa que el RMSE se reduce considerablemente con las distribuciones Half Cauchy y 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).

En el caso de $\sigma_1^2 = \sigma_2^2 = 3$ (**Figura 3-2**), de nuevo se observa que tanto para $\hat{\beta}$ como para $\hat{\gamma}$ el RMSE se reduce considerablemente con las distribuciones Half Cauchy y SBeta2 en comparación con la gamma-inversa, los resultados de la Half Cauchy y la SBeta2 son relativamente similares, aunque en muchos casos la SBeta2 presenta menores errores. Nuevamente se observan alternaciones entre crecimiento-decrecimiento en algunas de las

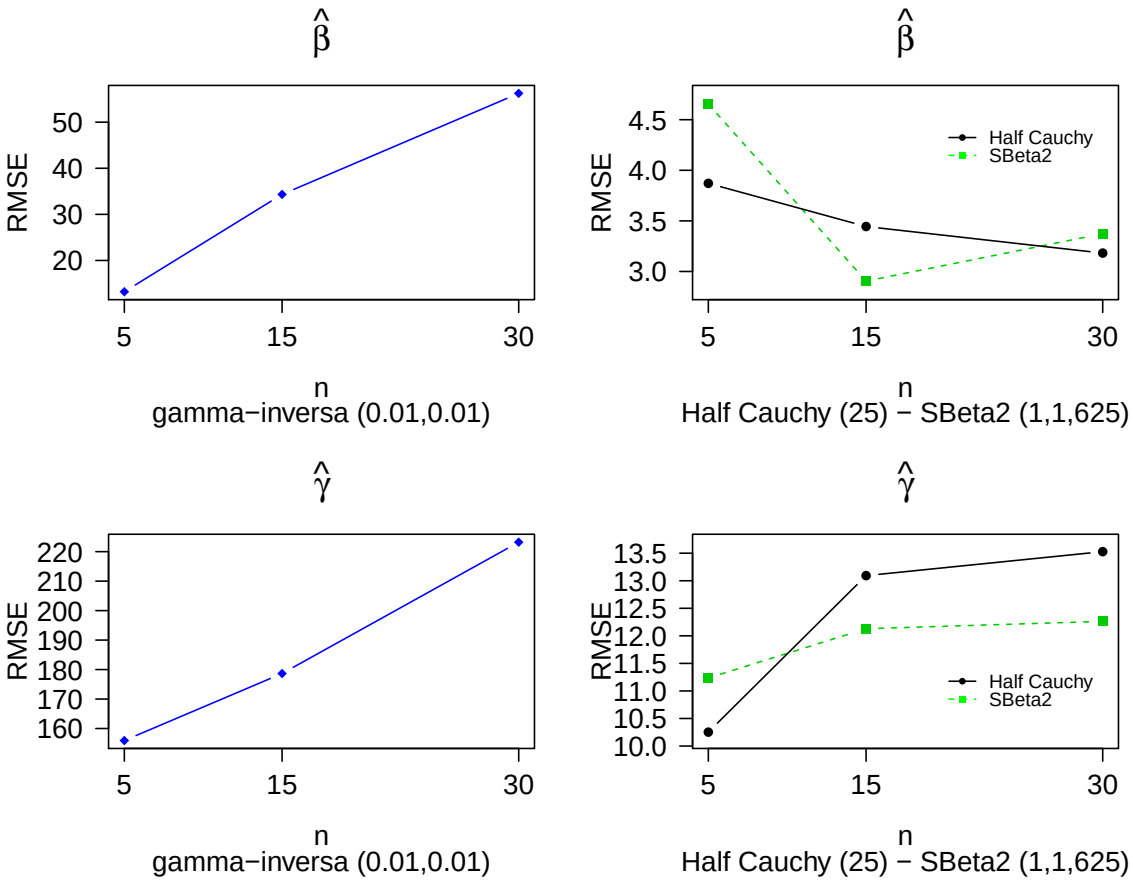

Figura 3-1.: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 0.1$

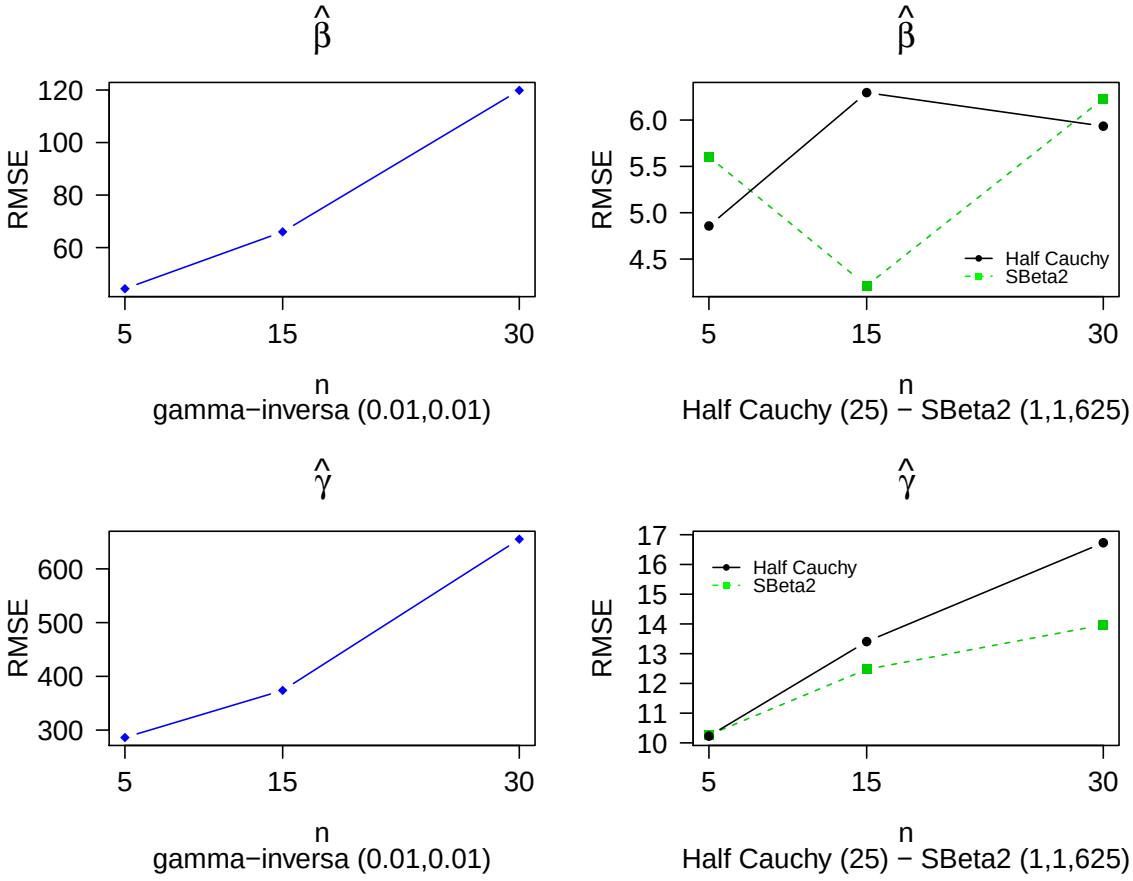

Figura 3-2.: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 3$

gráficas, o aumento del RMSE con el tamaño muestral, lo cual como ya se mencionó se explica como una resolución de conflicto (O'Hagan & Pericchi 2012).

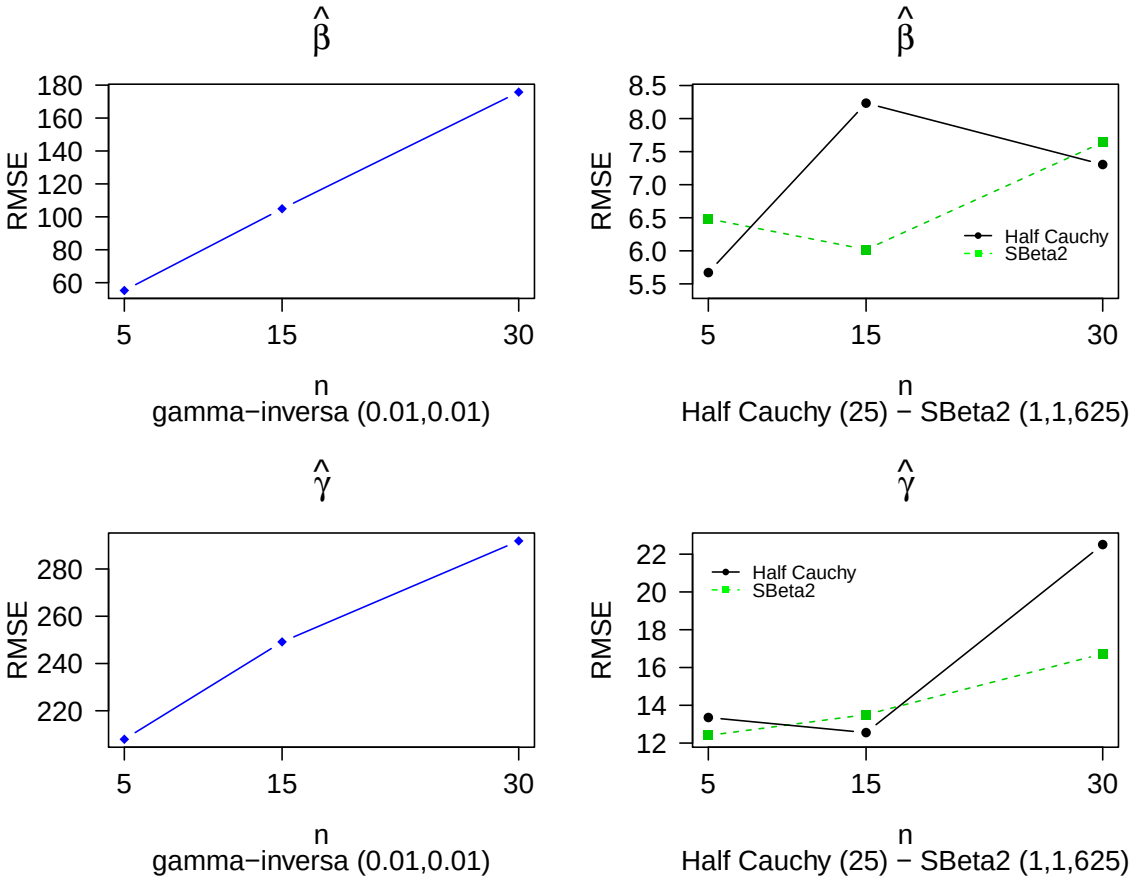

Figura 3-3.: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 10$

En el caso de $\sigma_1^2 = \sigma_2^2 = 10$ (**Figura 3-3**), en términos generales, los resultados son similares a los casos anteriores en cuanto a la reducción del RMSE con la distribuciones Half Cauchy y SBeta2 en comparación con la gamma-inversa, y que los resultados de la Half Cauchy y la SBeta2 son relativamente similares. También se evidencia la ocurrencia del fenómeno de resolución de conflicto (O'Hagan & Pericchi 2012).

En el caso de $\sigma_1^2 = \sigma_2^2 = 35$ (**Figura 3-4**), se confirman los resultados de los casos anteriores en cuanto a la reducción del RMSE con las distribuciones Half Cauchy y SBeta2 respecto a la gamma-inversa, que los resultados de la Half Cauchy y la SBeta2 son relativamente similares y la ocurrencia de una resolución de conflicto (O'Hagan & Pericchi 2012). Sin embargo, el hecho nuevo que debe recalcarse es que los resultados del caso $\sigma_1^2 = \sigma_2^2 = 35$ son muy similares a los del caso anterior ($\sigma_1^2 = \sigma_2^2 = 10$), lo que hace pensar que a este punto se

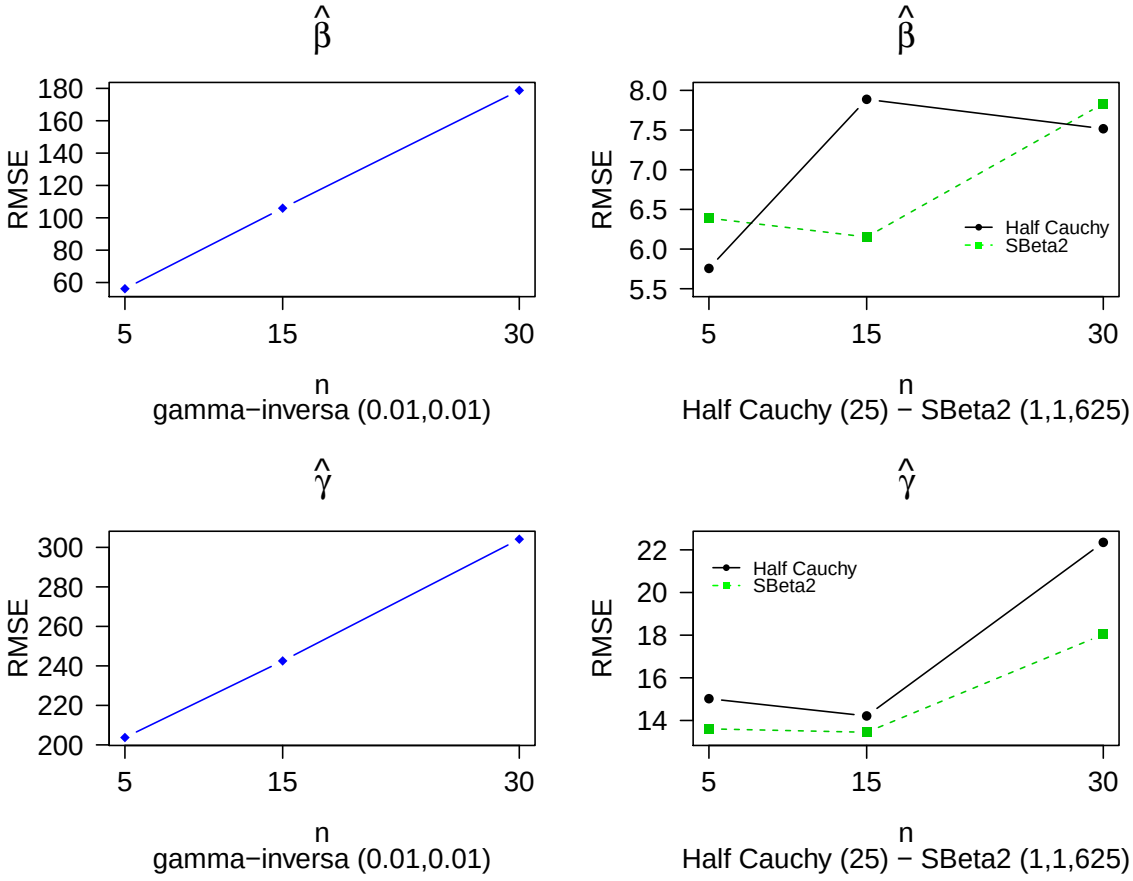

Figura 3-4.: RMSE VS Tamaño muestral - $\sigma_1^2 = \sigma_2^2 = 35$

da una estabilización del encogimiento, es decir, con cada una de las distribuciones candidatas el problema de encogimiento se estabiliza a partir de dichas condiciones de variabilidad.

Finalmente, como conclusión general del análisis de encogimiento para los diferentes valores de los parámetros de escala se tiene que dentro de las condiciones en que se enmarcó la simulación trabajada, 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.

3.1.1. Análisis de encogimiento ante la variación de los parámetros de escala

A continuación se presentará una re-visualización de los resultados ya obtenidos, comparando para cada una de las distribuciones candidatas el RMSE obtenido bajo un determinado tamaño muestral versus los diferentes valores de σ^2 (teniendo en cuenta que en cada uno de los escenarios trabajados σ_1^2 y σ_2^2 tomaban siempre el mismo valor).

De los gráficos que comparan el RMSE obtenido para cada una de las distribuciones candidatas versus los diferentes valores de σ^2 considerados y fijando $n = 5$ (**Figura 3-5**), primero se confirma que la distribución gamma-inversa presenta siempre los valores más altos de RMSE y que los resultados de la Half Cauchy y la SBeta2 son relativamente similares, aunque se tiene que para $\hat{\beta}$ resulta un poco mejor la Half Cauhy, pero para $\hat{\gamma}$ la distribución que ofrece mejores resultados es la SBeta2. De estos gráficos se confirma el resultado ya mencionado de la estabilización del encogimiento cuando σ^2 se hace grande, con cada una de las distribuciones candidatas (sin embargo, este comportamiento no es tan claro para las distribuciones Half Cauchy y SBeta2 con $\hat{\gamma}$).

En general, en cada gráfico se tiene un aumento del RMSE a medida que σ^2 crece, lo cual puede pensarse es un resultado intuitivo pues al aumentar la variabilidad introducida al modelo es más probable que los valores de los parámetros obtenidos vía simulación disten de los reales.

De los gráficos que comparan el RMSE obtenido para cada una de las distribuciones candidatas versus los diferentes valores de σ^2 considerados y fijando $n = 15$ (**Figura 3-6**), nuevamente se observa que la distribución gamma-inversa presenta siempre los valores más altos de RMSE y que los resultados de la Half Cauchy y la SBeta2 son relativamente similares. Sin embargo, bajo estas condiciones la distribución SBeta2 presenta mejores resultados tanto para $\hat{\beta}$ como para $\hat{\gamma}$. De nuevo se observa el aumento del RMSE cuando σ^2 crece y

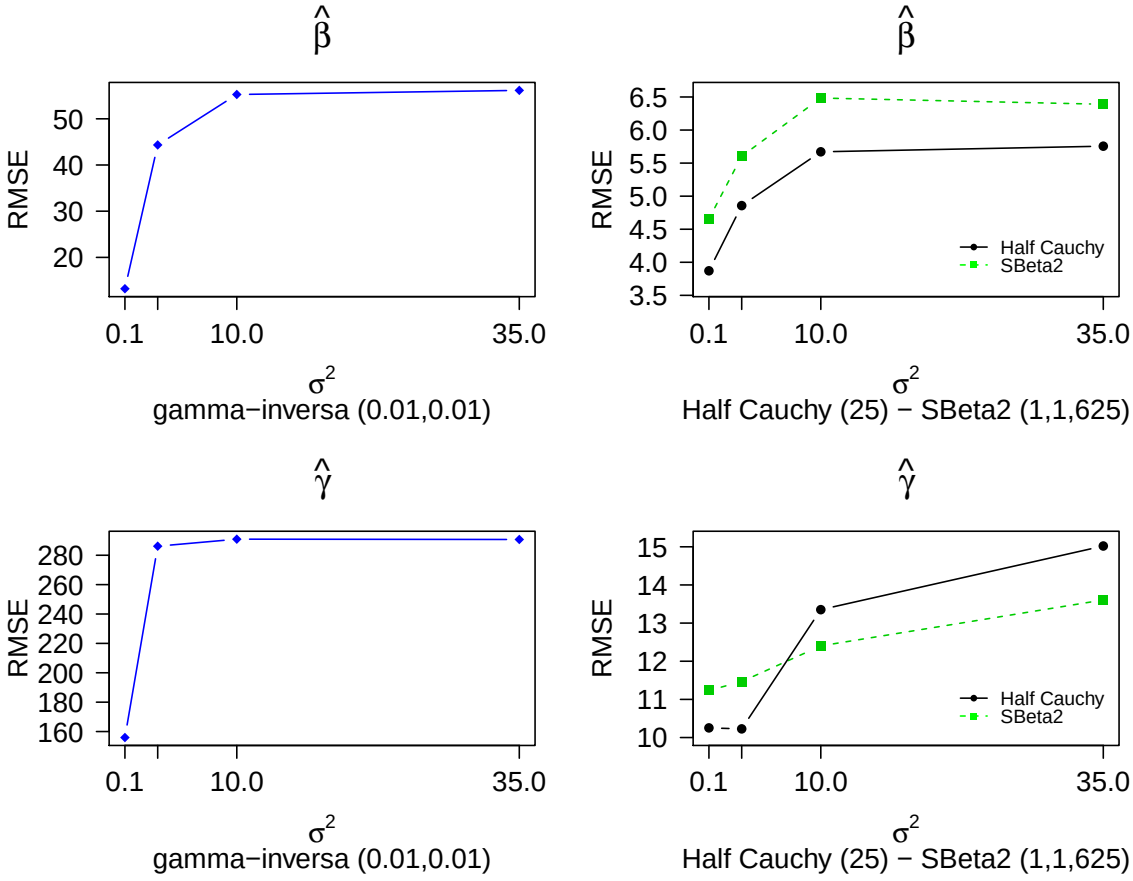

Figura 3-5.: RMSE VS σ^2 - $n = 5$

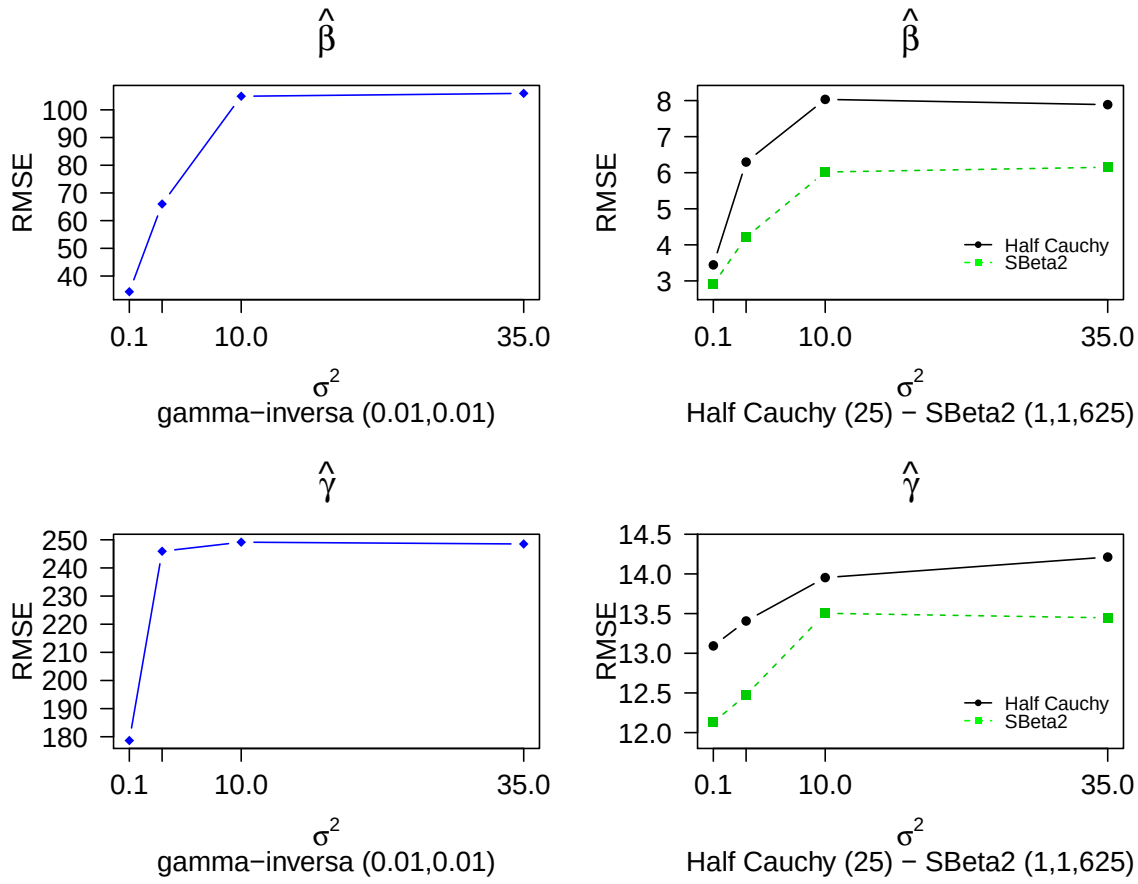

Figura 3-6.: RMSE VS σ^2 - $n = 15$

la estabilización del encogimiento también cuando σ^2 crece.

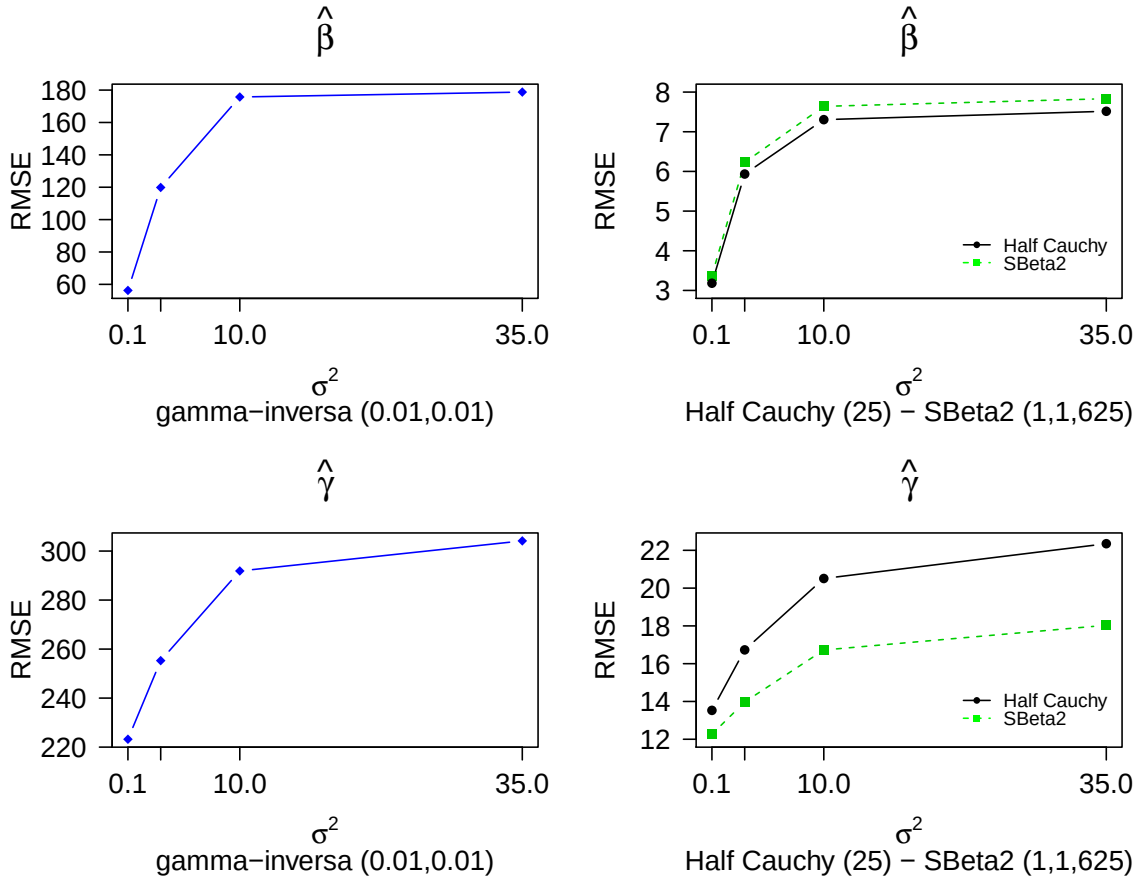

Figura 3-7.: RMSE VS σ^2 - $n = 30$

De los gráficos que comparan el RMSE obtenido para cada una de las distribuciones candidatas versus los diferentes valores de σ^2 considerados y fijando $n = 30$ (**Figura 3-7**), al igual que los casos anteriores se confirma que la distribución gamma-inversa presenta siempre los valores más altos de RMSE y que los resultados de la Half Cauchy y la SBeta2 son relativamente similares. Sin embargo, en este caso específicamente se tiene que para el parámetro $\hat{\gamma}$ la distribución SBeta2 presenta mejores resultados y para $\hat{\beta}$ los resultados de la Half Cauchy y la SBeta2 son muy cercanos. De nuevo se confirma el aumento del RMSE cuando σ^2 crece y la estabilización del encogimiento también cuando σ^2 crece.

3.1.2. 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 estaciona-

ria. Por esto se hace necesario realizar un chequeo de convergencia sobre las cadenas a posteriori obtenidas en este estudio de simulación. Es de mencionar que para las características de este trabajo, el chequeo de convergencia se hace muy extenso, pues se desarrollaron en total 36 escenarios (3 distribuciones candidatas, 4 valores asignados a los parámetros de escala, 3 tamaños muestrales) y la simulación de cada escenario se repitió 1000 veces, por tanto, se tienen 36000 cadenas para las que se debería realizar el chequeo de convergencia. De esta forma para verificar el supuesto de convergencia, lo que se hizo fue definir de forma aleatoria una cantidad moderada de cadenas y sobre estas realizar el chequeo. De forma ilustrativa se presenta el procedimiento desarrollado para una de las cadenas seleccionadas.

En 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 realizar un gráfico del *trace* o seguimiento de los valores generados del parámetro en cada una de la iteraciones de la cadena; además se evalúa 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).

A continuación 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 **figura 3-8** presenta el *trace* o seguimiento de los valores generados del parámetro, de dicho gráfico se observa que en general todos los valores generados fluctúan en un rango cercano. La **tabla 3-1** presenta los valores de la autocorrelación entre los valores generados al 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 3-1.: Autocorrelación - Cadena del análisis de encogimiento

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

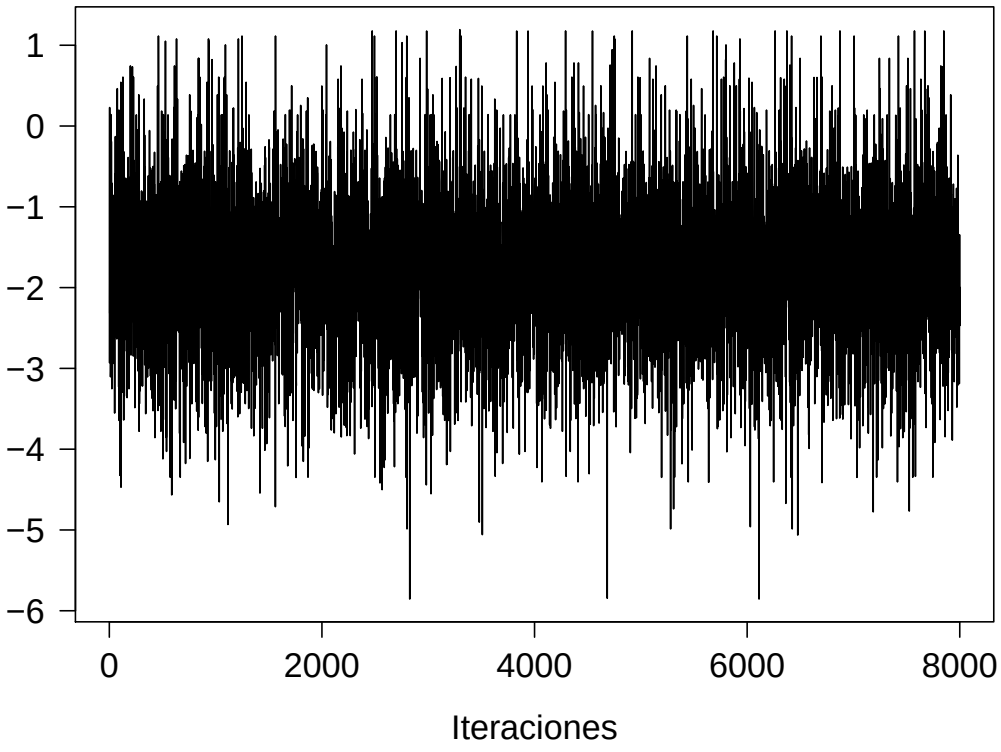

Figura 3-8.: Iteraciones - Cadena del análisis de encogimiento

La **figura 3-9** 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 observados en el gráfico de iteraciones, 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 en todas las demás cadenas en que se realizo el chequeo de convergencia estas cumplieron con el supuesto de alcanzar la distribución estacionaria.

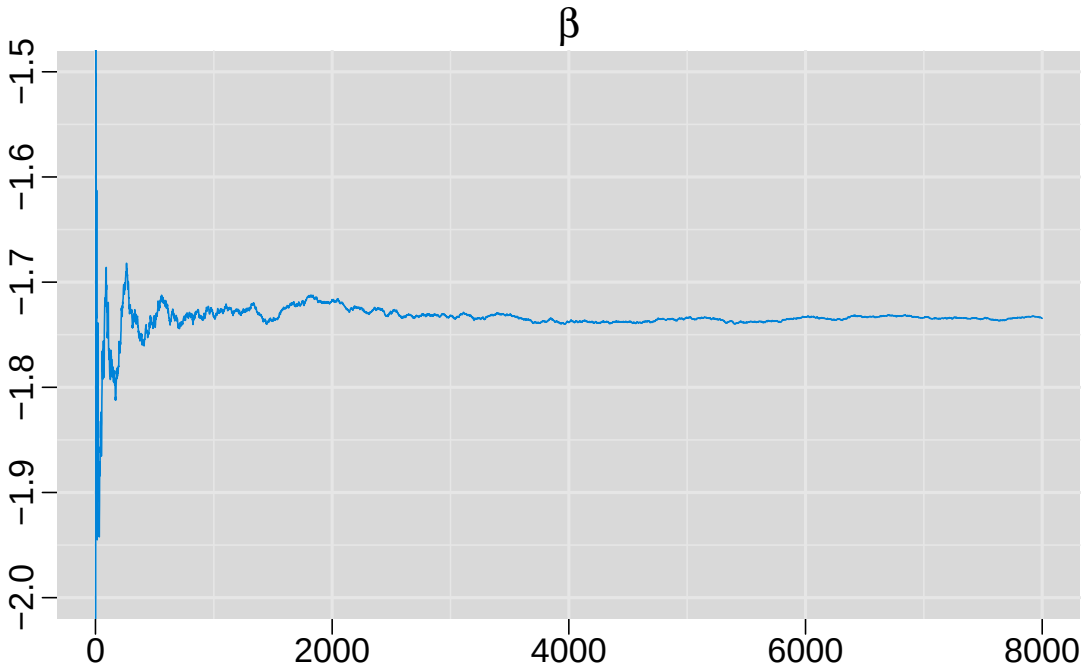

Figura 3-9.: Promedios móviles - Cadena del análisis de encogimiento

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 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 observaciones atípicas, siempre cada valor atípico se definió como dos veces el máximo valor de los datos originales de la variable respuesta del modelo. Este análisis se centró en dos escenarios principales: el caso en que se contaminaban los datos de la variable respuesta con una única observación atípica y el caso en que se contaminaban los datos de la variable respuesta con una cantidad moderada de observaciones atípicas (un tercio del total de datos).

El valor esperado de la variable respuesta del modelo se definió como el ajuste de las observaciones atípicas, 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 y escenario de observaciones atípicas (datos contaminados con un única observación atípica, o datos contaminados con una cantidad moderada de observaciones atípicas), nuevamente se generaron las cadenas a posteriori de los parámetros del modelo, cambiando los datos originales por los contaminados. Para cada parámetro del modelo, las cadenas a posteriori se construyeron de manera que contuvieran 10000 valores generados del parámetro, con un quemado inicial de 2000 valores, es decir, que finalmente cada cadena a posteriori contaba con 8000 valores generados del parámetro. 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 observaciones atípicas respecto al verdadero valor de las mismas para cada uno de los escenarios estudiados.

Para el caso en que los datos se contaminaron con una única observación atípica (**Figura 3-10**), 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. Finalmente se concluye que en el caso en que los datos presentan una única

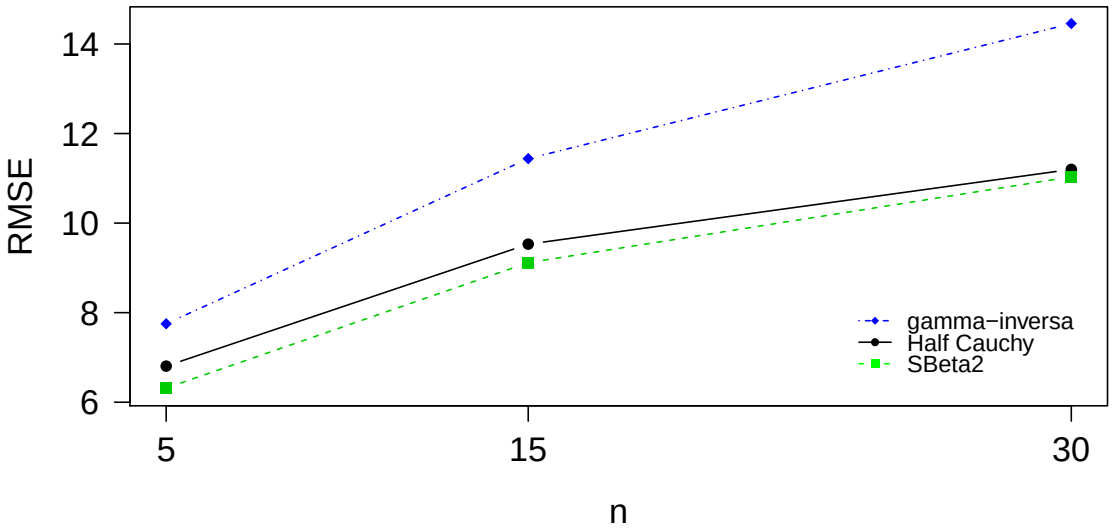

Figura 3-10.: RMSE VS Tamaño muestral - Ajuste con una única observación atípica

observación atípica, el modelo realiza un mejor ajuste del mismo si se usa la distribución SBeta2 como a priori para los parámetros de escala, con la distribución Half Cauchy el ajuste es relativamente similar, pero con la gamma-inversa se incurre en un mayor error de ajuste.

Para el caso en que los datos se contaminaron con una cantidad moderada de observaciones atípicas (un tercio del tamaño muestral) (**Figura 3-11**), nuevamente 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, lo cual puede explicarse como una resolución del conflicto (O'Hagan & Pericchi 2012). Igual que en el caso de una única observación atípica la distribución gamma-inversa presenta mayor error en el ajuste de las observaciones atípicas. Nuevamente los resultados para las distribuciones Half Cauchy y la SBeta2 son relativamente similares (mucho más cercanos que en el caso de una única observación atípica), aunque los errores en el ajuste de observaciones atípicas con la SBeta2 siempre son menores.

En general, puede observarse que en el caso en que los datos se contaminan con una cantidad moderada de observaciones atípicas bajo unos mismos escenarios los valores del RMSE son menores a los presentados en el caso de una única observación atípica, esto puede explicarse en cuanto que si al modelo se le introduce más datos contaminados, este estará en mayor capacidad de ajustar los mismos.

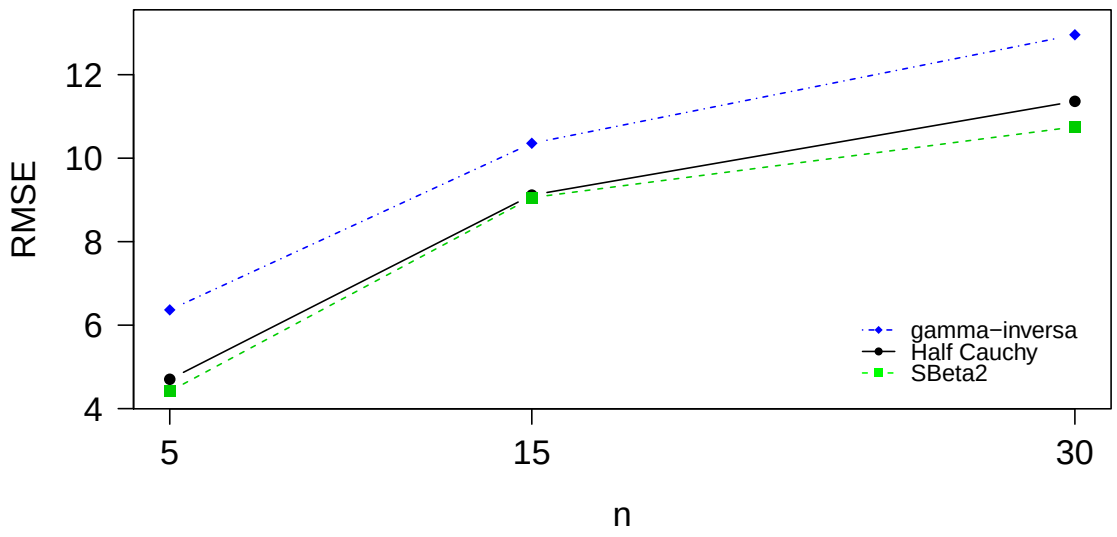

Figura 3-11.: RMSE VS Tamaño muestral - Ajuste con una cantidad moderada de observaciones atípicas

Finalmente, respecto al análisis de la capacidad del modelo para ajustar observaciones atípicas puede concluirse que dentro de las condiciones en que se enmarcó la simulación trabajada, la distribución gamma-inversa presenta mucha más dificultad a la hora de ajustar observaciones atípicas, ya sea que se tenga una única observación atípica o una cantidad moderada. 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.

3.2.1. Evaluación global del ajuste

Como complemento al análisis de la capacidad del modelo de ajustar observaciones atípicas, se evaluó el ajuste que el modelo con cada una de las distribuciones candidatas como a priori para los parámetros de escala realiza de cualquier observación (no sólo de las atípicas), y a partir de esto obtener una medida global del ajuste de observaciones que realiza el modelo con cada candidata. Para esto dentro de las condiciones en que los datos se

contaminaron con una única observación atípica, para cada uno de los tamaño muestrales considerados, con cada candidata se realizó el ajuste de dos observaciones no contaminadas, para cada observación se calcula el RMSE del ajuste de la observación respecto al verdadero valor de esta, y finalmente el promedio de los RMSE de las dos observaciones ordinarias y de la observación atípica se asumirá como la medida global del ajuste del modelo.

En la **tabla 3-2** aparecen los resultados del RMSE para cada una de las candidatas cuando $n = 5$, se tienen el RMSE del ajuste de dos observaciones no contaminadas (observaciones 1 y 3), de la observación atípica (observación 5) y la medida global del ajuste del modelo bajo cada candidata. De estos resultados se observa que los RMSE son muy similares con cada una de las candidatas para el ajuste de observaciones no contaminadas, la diferencia es un poco más marcada en el ajuste de la observación atípica y globalmente la distribución gamma-inversa presenta un mayor error en el ajuste de observaciones, con las distribuciones SBeta2 y Half Cauchy se obtienen mejoras, siendo los resultados de estas dos muy similares.

Tabla 3-2.: RMSE global de ajuste - $n = 5$

Observación	gamma inversa	SBeta2	Half Cauchy
1	0.9813	0.9785	0.9811
3	1.1922	1.2234	1.2155
5	7.7517	6.3158	6.8060
Global	3.3084	2.8392	3.0008

En la **tabla 3-3** aparecen los resultados del RMSE para cada una de las candidatas cuando $n = 15$, se tienen el RMSE del ajuste de dos observaciones no contaminadas (observaciones 1 y 8), de la observación atípica (observación 15) y la medida global del ajuste del modelo bajo cada candidata. De estos resultados nuevamente se observa que los RMSE son muy similares con cada una de las candidatas para el ajuste de observaciones no contaminadas, la diferencia es un poco más marcada en el ajuste de la observación atípica y de nuevo globalmente la distribución gamma-inversa presenta un mayor error en el ajuste de observaciones, con las distribuciones SBeta2 y Half Cauchy se obtienen mejoras, siendo los resultados de estas dos muy similares.

En la **tabla 3-4** aparecen los resultados del RMSE para cada una de las candidatas cuando $n = 30$, se tienen el RMSE del ajuste de dos observaciones no contaminadas (observaciones 1 y 16), de la observación atípica (observación 30) y la medida global del ajuste del modelo bajo cada candidata. De estos resultados nuevamente se observa que los RMSE son muy similares con cada una de las candidatas para el ajuste de observaciones no contaminadas, la diferencia es un poco más marcada en el ajuste de la observación atípica y de nuevo

Tabla 3-3.: RMSE global de ajuste - $n = 15$

Observación	gamma inversa	SBeta2	Half Cauchy
1	1.5812	1.2560	1.3424
8	2.1693	2.2159	2.1835
15	11.4400	9.1165	9.5300
Global	5.0635	4.1961	4.3520

globalmente la distribución gamma-inversa presenta un mayor error en el ajuste de observaciones, con las distribuciones SBeta2 y Half Cauchy se obtienen mejoras, siendo los resultados de estas dos muy similares.

Tabla 3-4.: RMSE global de ajuste - $n = 30$

Observación	gamma inversa	SBeta2	Half Cauchy
1	2.0219	2.0196	2.0179
16	2.8281	2.8677	2.6684
30	14.4562	11.0332	11.1991
Global	6.4354	5.3068	5.2951

Finalmente, del análisis global del ajuste del modelo de observaciones bajo cada distribución candidata como a priori para los parámetros de escala, se tiene que el ajuste de observaciones no contaminadas es muy similar con todas las candidatas, pero cuando se ajustan observaciones atípicas sí se presenta mayor diferencia. En términos globales bajo la distribución gamma-inversa se tiene un mayor error cuando se ajustan observaciones, con las distribuciones SBeta2 y Half Cauchy los ajustes mejoran y se obtienen resultados relativamente similares.

3.2.2. Chequeo de convergencia

El chequeo de convergencia para las cadenas a posteriori obtenidas en el análisis de la capacidad del modelo de ajustar observaciones atípicas se realizó de la misma forma como se hizo para el análisis de encogimiento. Así, se escoge aleatoriamente un número moderado de cadenas entre todas las simuladas y sobre estas se realizó el chequeo. A continuación se presentan los resultados del chequeo de convergencia para la cadena obtenida bajo las condiciones: cantidad moderada de observaciones atípicas, distribución candidata gamma-inversa, $n = 30$, simulación número 427 del parámetro γ.

La **figura 3-12** presenta el *trace* o seguimiento de los valores generados del parámetro, de dicho gráfico se observa que en general todos los valores generados fluctúan en un rango cercano. La **tabla 3-5** 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.

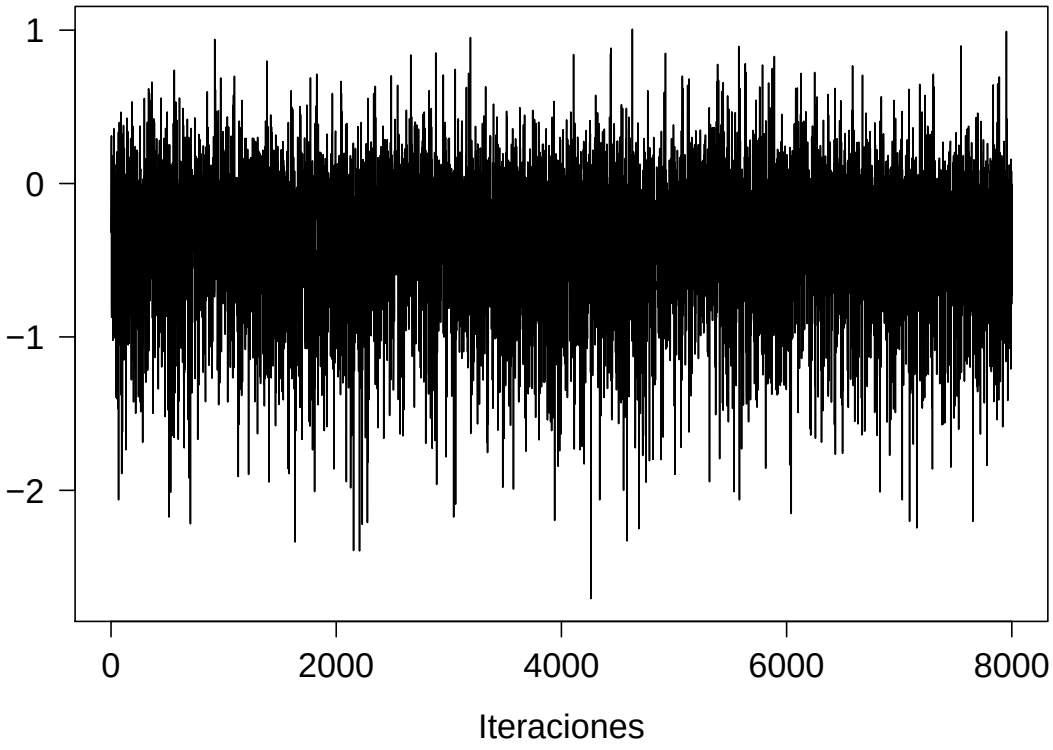

Figura 3-12.: Iteraciones - Cadena del análisis del ajuste de observaciones atípicas

Tabla 3-5.: Autocorrelación - Cadena del análisis del ajuste de observaciones atípicas

	1 rezago	5 rezagos	10 rezagos	50 rezagos
γ	0.020160078	-0.005266224	0.021227197	0.027091117

La **figura 3-13** 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 0.0704 y valor p de 0.1, con lo cual se concluye que no existe suficiente evidencia muestral para rechazar la hipótesis nula de que la cadena alcanza la distribución estacionaria. Así, dados los resultados observados en el gráfico de iteraciones, 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 en todas las demás cadenas del análisis de la capacidad del modelo de ajustar observaciones atípicas en que se realizó el chequeo de convergencia estas cumplieron con el supuesto de alcanzar la distribución estacionaria.

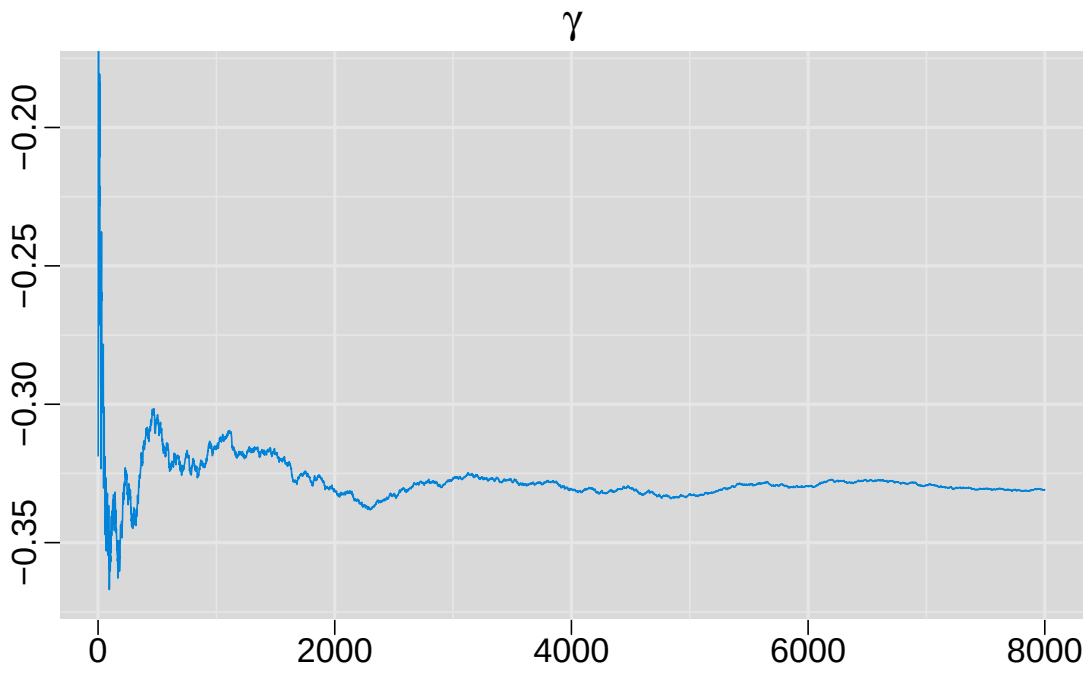

Figura 3-13.: Promedios móviles - Cadena del análisis del ajuste de observaciones atípicas

4. Caso práctico

Como ilustración de la metodología propuesta en el presente trabajo, se realizó una aplicación de la misma con datos obtenidos de un estudio de horticultura, es de mencionar que la horticultura es la ciencia encargada de estudiar los métodos de cultivo de hortalizas. Específicamente los datos a trabajar hacen referencia al cultivo de manzanas los cuales fueron obtenidos por Marin et al. (1993) y se presentan en la **tabla 4-1**. Los datos son el número de raíces producidas por 270 brotes micropropagados de la columna de cultivos de manzana tipo Trajan. Durante el período de enraizamiento todos los brotes se mantuvieron en condiciones idénticas, sin embargo, la diferencia entre dichos brotes es que estaban cultivados en medios que contenían diferentes concentraciones de cytokinin BAP, la cual es una proteína utilizada para mejorar las condiciones del suelo, y además los brotes crecían en cámaras de cultivo expuestas a condiciones de fotoperiodo (iluminación proporcionada a la planta) de 8 y 16 horas. De esta forma se tiene que dichos 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.

En la **tabla 4-2** se presenta una comparación del ajuste de los datos de cultivo de manzanas al modelo de regresión Poisson y al modelo ZIP, las comparaciones se discriminan por fotoperiodo (8 y 16 horas), en dicha tabla O hace referencia al número de observaciones de una determinada cantidad de raíces, E_{ZIP} hace referencia al valor esperado de observaciones para una determinada cantidad de raíces si se asume que dichos valores distribuyen ZIP y E_P hace referencia al valor esperado de observaciones para una determinada cantidad de raíces si se asume que dichos valores distribuyen Poisson. Estos resultados fueron tomados de (Rodrigues 2006), de los cuales se tiene que en el caso de un fotoperiodo de 8 horas, los valores esperados asumiendo que los datos distribuyen ZIP son cercanos a los valores esperados cuando se asume que distribuyen Poisson, y a la vez cercanos a los valores observados. Sin embargo, bajo un fotoperiodo de 16 horas, existe una diferencia marcada entre los valores esperados bajo la distribución ZIP y la Poisson, siendo los valores esperados ZIP más cercanos a los observados. En (Rodrigues 2006), además se realiza la prueba chi-cuadrado discriminada por fotoperiodo, para evaluar la bondad de ajuste de la distribución ZIP y la Poisson, encontrándose que los datos de cultivos de manzanas obtenidos por Marin et al. (1993) se ajustan mucho mejor a la distribución ZIP.

Tabla 4-1.: Datos cultivo de manzanas (Marin et al. 1993)

	Fotoperiodo							
	8				16			
BAP	2.2	4.4	8.8	17.6	2.2	4.4	8.8	17.2
Raíces								
0	0	0	0	2	15	16	12	19
1	3	0	0	0	0	2	3	2
2	2	3	1	0	2	1	2	2
3	3	0	2	2	2	1	1	4
4	6	1	4	2	1	2	2	3
5	3	0	4	5	2	1	2	1
6	2	3	4	5	1	2	3	4
7	2	7	4	4	0	0	1	3
8	3	3	7	8	1	1	0	0
9	1	5	5	3	3	0	2	2
10	2	3	4	4	1	3	0	0
11	1	4	1	4	1	0	1	0
12	0	0	2	0	1	1	1	0
13	1	1	0	0	0	0	0	0
14	0	0	2	1	0	0	0	0
17	1	0	0	0	0	0	0	0

Tabla 4-2.: Comparación del ajuste de datos cultivo manzanas - Tomado de (Rodrigues 2006)

	Fotoperiodo					
	8			16		
	O	E_{ZIP}	E_P	O	E_{ZIP}	E_P
Raíces						
0	2	2.91	0.11	62	61.68	7.43
1	3	0.78	0.82	7	1.74	21.27
2	6	2.78	2.91	7	4.63	30.43
3	7	6.6	6.89	8	8.23	29.02
4	13	11.75	12.23	8	11	20.76
5	12	16.76	17.36	6	11.8	11.88
6	14	19.93	20.55	10	10.57	5.66
7	17	20.34	20.84	4	8.13	2.31
8	21	18.18	18.5	2	5.49	0.82
9	14	14.45	14.59	7	3.31	0.26
10	13	10.35	10.36	4	1.79	0.07
11	10	6.74	6.68	2	0.89	0.01
12	2	4.03	3.95	3	0.4	0
13	2	2.2	2.16			
14	3	1.14	1.09			
17	1	0.54	0.51			

Dadas las condiciones y resultados mencionados anteriormente, se tiene que es pertinente trabajar bajo el modelo ZIP, con la siguiente estructura:

$$\begin{aligned} &Y_i \sim ZIP(p_i, \lambda_i), \\ &log(\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \\ &logit(p_i) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2. \end{aligned} \tag{4-1}$$

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 en el medio. Además, se asume que $\beta_0 \sim U(-2.5, 2.5)$, $\beta_1 \sim N(0, \sigma_1^2)$, $\beta_2 \sim N(0, \sigma_2^2)$, $\gamma_0 \sim U(-2.5, 2.5)$, $\gamma_1 \sim N(0, \sigma_3^2)$, $\gamma_2 \sim N(0, \sigma_4^2)$.

A continuación se procedió a realizar el ajuste del modelo ZIP bajo las condiciones planteadas en 4-1, definiendo como a priori para los parámetros de escala cada una de las distribuciones candidatas del presente trabajo (una a la vez). Después con cada una de las candidatas se procedió a simular las cadenas a posteriori de los parámetros del modelo, por medio del método MCMC, dicho procedimiento se realizó en el software OpenBUGS®, al

cual para cada distribución candidata se le introdujo el modelo ajustado y los datos presentados en la **tabla 4-1**, para cada una de los parámetros se generaron 100000 valores por cadena, con un quemado inicial de 20000 valores, es decir, que finalmente cada cadena a posteriori dispone de 80000 valores generados del parámetro.

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 presentados en la **tabla 4-1** con una observación atípica y observar bajo cuál distribución candidata se realiza un mejor ajuste de la misma. No se realizó un análisis de encogimiento de los parámetros a posteriori del modelo, pues para dicha comparación es necesario conocer los verdaderos valores de los parámetros.

En la **tabla 4-3** 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 (número de parámetros a considerar), 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. Para el cálculo del DIC se parte del valor de la *Deviance* la cual se obtiene a partir de la función de verosimilitud.

Tabla 4-3.: Medida de ajuste - Distribuciones candidatas

Distribución	DIC
gamma-inversa	1873
SBeta2	1870
Half Cauchy	1871

De los resultados presentados en la **tabla 4-3** 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.

En la **tabla 4-4** 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. En general, puede observarse 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 sin importan la a priori de la que se parta, las distribuciones a posteriori van a estar más influenciadas por la información muestral, lo cual genera estimaciones a posteriori cercanas entre las candidatas.

Tabla 4-4.: 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

Como ya se había mencionado anteriormente, dentro del análisis de este caso práctico también se realizó una evaluación de la capacidad del modelo de ajustar una observación atípica, bajo cada distribución candidata asignada como a priori para los parámetros de escala. Para esto se tomó un valor de la variable respuesta (número de raíces en los brotes) y se sustituyó por dos veces el máximo valor de los datos originales, obteniéndose así una observación atípica y posteriormente con cada distribución candidata se realizó un ajuste de dicha observación, el ajuste se obtuvo por medio del valor esperado de la variable ZIP, sabiendo que si $Y \sim ZIP(p, \lambda)$, entonces $E(Y) = (1-p)\lambda$. De esta forma, lo que se hizo fue a partir de las cadenas a posteriori de los parámetros del modelo se calcula el valor esperado de la variable respuesta, el cual se asumió como el valor que ajusta la observación atípica. La **tabla 4-5** presenta los resultados del ajuste de la observación atípica con cada una de las distribuciones candidatas.

Tabla 4-5.: 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

De los resultados presentados en la **tabla 4-5** 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.

Finalmente, de forma general, de los resultados del caso práctico presentado 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.

4.1. Chequeo de convergencia

Como se mencionó en el estudio de simulación (en la sección 3.1.2), sobre las cadenas a posteriori es necesario realizar un chequeo de convergencia, para comprobar que estas cumplen con el supuesto de alcanzar la distribución estacionaria. El chequeo de convergencia se realizó sobre todas las cadenas obtenidas en este caso práctico, de manera ilustrativa se presentan los resultados del chequeo de convergencia para la cadena obtenida bajo las condiciones: distribución candidata Half Cauchy, parámetro γ_2, datos originales, es decir, sin la presencia de una observación atípica.

La **figura 4-1** presenta el *trace* o seguimiento de los valores generados del parámetro, de dicho gráfico se observa que en general todos los valores generados fluctúan en un rango cercano. La **tabla 4-6** 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 4-6.: Autocorrelación - Caso práctico

	1 rezago	5 rezagos	10 rezagos	50 rezagos
γ_2	0.002713980	-0.003970850	0.002279277	0.005879868

La **figura 4-2** 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 0.1734 y valor p de 0.1, con lo cual se concluye que no existe suficiente evidencia muestral para rechazar la hipótesis nula de que la cadena alcanza la distribución estacionaria. Así, dados los resultados observados en el gráfico de iteraciones, 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 cadenas a posteriori obtenidas en el desarrollo del caso práctico cumplen con el supuesto de alcanzar la distribución estacionaria.

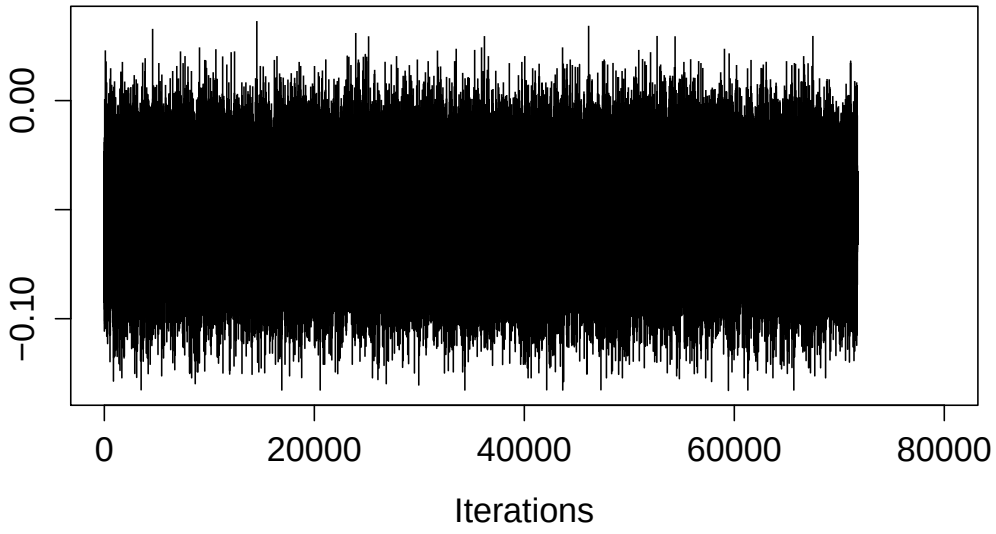

Figura 4-1.: Iteraciones - Cadena del caso práctico

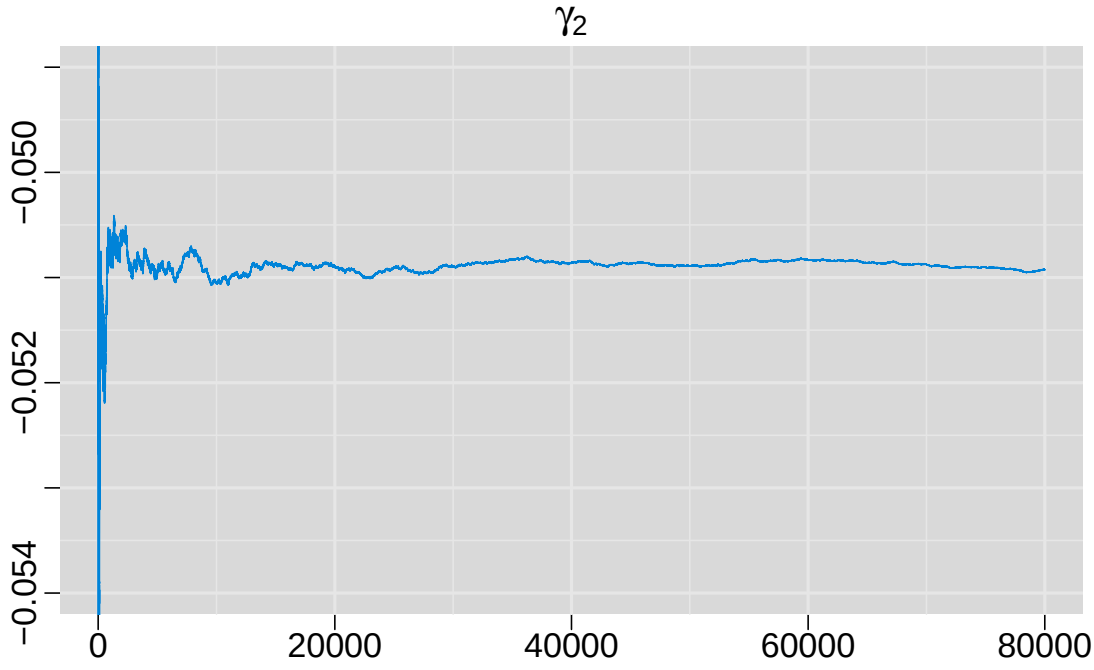

Figura 4-2.: Promedios móviles - Cadena del caso práctico

5. Conclusiones y trabajo futuro

5.1. Conclusiones

A continuación se presentan las principales conclusiones obtenidas de todo el proceso de investigación (revisión bibliográfica, estudio de simulación y caso práctico):

- La distribución gamma-inversa ha sido ampliamente utilizada como a priori para los parámetros de escala, sin embargo, existen fuertes críticas a esta práctica, por ejemplo, Berger (2006) plantea que al hacerlo se obtiene una distribución a posteriori sesgada en valores cercanos a cero, lo cual desmejora el ajuste de observaciones atípicas; Gelman (2006) argumenta que con la gamma-inversa(ϵ, ϵ) buscando que sea no informativa se hace $\epsilon \to 0$, lo cual en realidad produce un encogimiento en los parámetros a posteriori.

- La distribución Half Cauchy propuesta por Gelman (2006) como a priori para los parámetros de escala en modelos jerárquicos, desde el punto de vista teórico ofrece ventajas en dicho uso, 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.

- La distribución SBeta2, para la cual Pérez et al. (2015) estudian las ventajas de usarla como a priori para parámetros de escala, desde el punto de vista teórico ofrece ventajas en dicho uso, pues si la varianza distribuye SBeta2, entonces la precisión también; Valores de la SBeta2 son fácilmente simulables, la distribución puede integrarse al esquema del muestreador de Gibbs; Es una distribución flexible, capaz de modelar diferentes tipos de comportamientos en el origen y la cola; La SBeta2 es una distribución robusta, donde el espesor de su cola es equivalente al de la t-student (Pérez et al. 2015).

- Dentro de las condiciones en que se enmarcó el estudio de simulación desarrollado en este trabajo, 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 presentado 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.

5.2. Trabajo futuro

Algunos proyectos de investigación que se considera pueden desarrollarse a partir del presente trabajo son:

- Realizar un análisis similar de distribuciones a priori para los parámetros de escala, en otros modelos utilizados en fenómenos de conteo con presencia excesiva de ceros. Por ejemplo, se podría trabajar sobre el modelo Binomial Negativo inflado con ceros.

- Sobre el mismo modelo ZIP realizar un análisis conjunto de distribuciones a priori para los parámetros de localización y escala.

- Definir de manera precisa qué distribución representa una mejor a priori para los parámetros de escala del modelo ZIP, si la SBeta2 o la Half Cauchy. Para esto puede usarse por ejemplo el factor de Bayes.

- Considerar el caso en que exista dependencia entre los parámetros del modelo ZIP, y bajo dicha situación definir una a priori multivariable.

A. Anexo

A.1. Códigos de programación - Estudio de simulación

A.1.1. Análisis de encogimiento

```
### Generación de datos ZIP ###
## m es el número de muestras ZIP que se van a generar
m <- 1000
## Fijando el valor de los parámetros de escala
sigma1<- 0.1
sigma2<- 0.1
set.seed(123)
nor <- rnorm(2*m,0,sigma1)
beta1 <- nor[1:1000]
gamma1 <- nor[1001:2000]
set.seed(123)
int = runif(m,-2.5,2.5)
set.seed(123)
x_data <- runif(m,0,1)
lambda_data <- exp(beta1*x_data)
library(boot)
p_data <- inv.logit(int + gamma1*x_data)
library(gamlss.dist)
n <- 5 # cantidad de datos ZIP a generar
S <- rep(n,m)
M<- data.frame(S,lambda_data,p_data)
y_data <- t(mapply(rZIP,n=M$S,mu=M$lambda_data,sigma=M$p_data))

#### Ajuste del modelo ZIP - candidata gamma-inversa ####
library(coda)
library(lattice)
library(R2WinBUGS)
library(R2OpenBUGS)

zip.model<- function(){
  C<-0
```

```
  for(i in 1:N){
    zeros[i] <- 0
    zeros[i] ~ dpois( zeros.means[i] )
    zeros.means[i] <- -l[i] + C
    # log-likelihood for i individual
    l[i] <- log( p[i] * equals( y[i], 0 ) + (1-p[i])*fd[i] )
    # set the probability function of each distribution
    fd[i] <-  exp( -lambda.ind[i] + y[i]*log(lambda.ind[i]) - loggam(y[i]+1) )
    log(lambda.ind[i]) <- beta * x[i]
    logit(p[i]) <- int + gamma*x[i]
  }
  lambda<- exp(beta)
  int~dunif(-2.5,2.5)
  beta~dnorm(0,tau1)
  gamma~dnorm(0,tau2)
  tau1~dgamma(0.01,0.01)
  tau2~dgamma(0.01,0.01)
  sigma1 <- 1/tau1
  sigma2 <- 1/tau2
}

#### Ajuste del modelo ZIP - candidata SBeta2 ####
zip.model<- function(){
  C<-0
  for(i in 1:N){
    zeros[i] <- 0
    zeros[i] ~ dpois( zeros.means[i] )
    zeros.means[i] <- -l[i] + C
    # log-likelihood for i individual
    l[i] <- log( p[i] * equals( y[i], 0 ) + (1-p[i])*fd[i] )
    # set the probability function of each distribution
    fd[i] <-  exp( -lambda.ind[i] + y[i]*log(lambda.ind[i]) - loggam(y[i]+1) )
    log(lambda.ind[i]) <- beta * x[i]
    logit(p[i]) <- int + gamma*x[i]
  }
  lambda<- exp(beta)
  int~dunif(-2.5,2.5)
  beta~dnorm(0,tau1)
  gamma~dnorm(0,tau2)
  sigma1 <- b*(H1/(1-H1))
  sigma2 <- b*(H2/(1-H2))
  b <- 625
```

```
  H1~dbeta(1,1)
  H2~dbeta(1,1)
  tau1 <- 1/sigma1
  tau2 <- 1/sigma2
}

#### Ajuste del modelo ZIP - candidata Half-Cauchy ####
zip.model<- function(){
  C<-0
  for(i in 1:N){
    zeros[i] <- 0
    zeros[i] ~ dpois( zeros.means[i] )
    zeros.means[i] <- -l[i] + C
    # log-likelihood for i individual
    l[i] <- log( p[i] * equals( y[i], 0 ) + (1-p[i])*fd[i] )
    # set the probability function of each distribution
    fd[i] <-  exp( -lambda.ind[i] + y[i]*log(lambda.ind[i]) - loggam(y[i]+1) )
    log(lambda.ind[i]) <- beta * x[i]
    logit(p[i]) <- int + gamma*x[i]
  }
  lambda<- exp(beta)
  int~dunif(-2.5,2.5)
  beta~dnorm(0,tau1)
  gamma~dnorm(0,tau2)
  sigma1 <- b*(H1/(1-H1))
  sigma2 <- b*(H2/(1-H2))
  b <- 625
  H1~dbeta(0.5,0.5)
  H2~dbeta(0.5,0.5)
  tau1 <- 1/sigma1
  tau2 <- 1/sigma2
}

### Función que permite la simulación MCMC múltiple de un mismo escenario ###
sim <- function(y,x,N)
{
  zip.data <- list("y","x","N")
  zip.params <- c("sigma1","sigma2","beta","gamma")
  #Valores iniciales
  zip.inits <- function(){
    list("beta"=c(0.01), "gamma"=c(0.01),"tau1"=c(10),"tau2"=c(10))
  }
```

```
  zipfit <- bugs(data=zip.data, inits=zip.inits, parameters=zip.params, DIC=TRUE,
model.file = zip.model, n.chains = 1, n.iter = 10000, n.burnin = 2000, n.thin = 1)
  attach.bugs(zipfit,overwrite = TRUE)
  t=rbind(beta,gamma,deviance)
  return(list(a=t,dic=zipfit$DIC))
  detach.bugs(zipfit)
}

## Introducción de datos
y <- matrix(y_data,ncol=n)
set.seed(123)
x <- runif(n,0,1)
N <- n
# Ejecutar la simulación MCMC
f <- apply(y,1,sim,x=x,N=N)
g=unlist(f,use.names = FALSE)
k=24001 # 3*(N° cadenas) + 1
dic <-NULL
for (i in 1:m){
  dic[i]=g[k]
  k=k+24001
}
#l=g[-(seq(3*8000+1,3*8000*m+m,3*8000+1))] # a g se le extrae el dic
l=g[-(seq(24001,24001000,24001))]
r=matrix(l,ncol=m,nrow=24000,byrow=FALSE) #nrow=3*(N° cadenas)
beta=r[seq(1,24000,3),]
gamma=r[seq(2,24000,3),]
deviance=r[seq(3,24000,3),]

## Análisis RMSE (Raíz error cuadrático medio)
mb=apply(beta,2,median)
b=(beta1-mb)^2
(rmse_b=sqrt(sum(b)/1000))
mg=apply(gamma,2,median)
g=(gamma1-mg)^2
(rmse_g=sqrt(sum(g)/1000))
```

A.1.2. Análisis - Ajuste de observaciones atípicas

```
# Generación de un outlier
ma <- max(y_data)
for (i in 1:m){
  y_data[i,n]=2*ma
```

```
}
# Generación de diez outliers
ma <- max(y_data)
for (i in 1:m){
  for (j in 0:9)
    y_data[i,n-j]=2*ma
}
# Valor verdadero del outlier
(OL <- 2*max(y_data) )
# Beta y Gamma posteriores
beta <- read.table(file = "D:/Usuario/Desktop/Resultados/Outliers/HC_n5_pocos_beta.txt")
gamma <- read.table(file = "D:/Usuario/Desktop/Resultados/Outliers/HC_n5_pocos_gamma.txt")
# Datos de la covariable
set.seed(123)
x <- runif(n,0,1)
# Medianas de Beta y Gamma
b <- apply(beta,2,median)
g <- apply(gamma,2,median)
# P y Lambda posteriores
lambda <- exp(b*x[n]) # x[n] es la covariable que corresponde al outlier
p <- inv.logit(int + g*x[n])
# Esperanza de Y, que se usará para ajustar el outlier
E_Y <- (1-p)*lambda
# RMSE del ajuste del outlier
l <- (E_Y-OL)^2
(rmse <- sqrt(sum(l)/1000))
```

A.1.3. Chequeo de convergencia

```
### Análisis de encogimiento ###
## Cadenas posteriores obtenidas en el estudio de simulación
b <- read.table(file = "D:/Usuario/Desktop/Resultados/SB_n15_1s01_2s01_beta.txt")
## Se trabajan con las cadenas de una única simulación
beta <- b[1:8000,83]
## Convertir las cadenas en un objeto MCMC
library(coda)
beta.mcmc <- mcmc(beta)
#  Gráfico del Trace
traceplot(beta.mcmc)
# Valores autocorrelación
autocorr.diag(beta.mcmc)
# Gráfico medias moviles
library(mcmcplots)
```

```
rmeanplot(beta.mcmc , main =expression(beta))
## Test kpss - Hipótesis convergencia
library(tseries)
kpss.test(beta)
```

A.2. Códigos de programación - Caso práctico

```
### Modelo ZIP - Ajuste gamma-inversa ###
zip.model<- function(){
  C<-0
  for(i in 1:N){
    zeros[i] <- 0
    zeros[i] ~ dpois( zeros.means[i] )
    zeros.means[i] <- -l[i] + C
    # log-likelihood for i individual
    l[i] <- log( p[i] * equals( y[i], 0 ) + (1-p[i])*fd[i] )
    # set the probability function of each distribution
    fd[i] <-  exp( -lambda.ind[i] + y[i]*log(lambda.ind[i]) - loggam(y[i]+1) )
    log(lambda.ind[i]) <- b1 + beta1*x1[i] + beta2*x2[i]
    logit(p[i]) <- b0 + gamma1*x1[i] + gamma2*x2[i]
  }
  lambda<- exp(beta1 + beta2)
  b0~dunif(-2.5,2.5)
  beta1~dnorm(0,tau1)
  gamma1~dnorm(0,tau2)
  beta2~dnorm(0,tau2)
  gamma2~dnorm(0,tau1)
  tau1~dgamma(0.01,0.01)
  tau2~dgamma(0.01,0.01)
  sigma1 <- 1/tau1
  sigma2 <- 1/tau2
}

### Modelo ZIP - Ajuste SBeta2 ###
zip.model<- function(){
  C<-0
  for(i in 1:N){
    zeros[i] <- 0
    zeros[i] ~ dpois( zeros.means[i] )
    zeros.means[i] <- -l[i] + C
    # log-likelihood for i individual
    l[i] <- log( p[i] * equals( y[i], 0 ) + (1-p[i])*fd[i] )
```

```
    # set the probability function of each distribution
    fd[i] <-  exp( -lambda.ind[i] + y[i]*log(lambda.ind[i]) - loggam(y[i]+1) )
    log(lambda.ind[i]) <- b1 + beta1*x1[i] + beta2*x2[i]
    logit(p[i]) <- b0 + gamma1*x1[i] + gamma2*x2[i]
  }
  lambda<- exp(beta1 + beta2)
  b0~dunif(-2.5,2.5)
  beta1~dnorm(0,tau1)
  gamma1~dnorm(0,tau2)
  beta2~dnorm(0,tau2)
  gamma2~dnorm(0,tau1)
  sigma1 <- b*(H1/(1-H1))
  sigma2 <- b*(H2/(1-H2))
  b <- 625
  H1~dbeta(1,1)
  H2~dbeta(1,1)
  tau1 <- 1/sigma1
  tau2 <- 1/sigma2
}

### Modelo ZIP - Ajuste Half Cauchy ###
zip.model<- function(){
  C<-0
  for(i in 1:N){
    zeros[i] <- 0
    zeros[i] ~ dpois( zeros.means[i] )
    zeros.means[i] <- -l[i] + C
    # log-likelihood for i individual
    l[i] <- log( p[i] * equals( y[i], 0 ) + (1-p[i])*fd[i] )
    # set the probability function of each distribution
    fd[i] <-  exp( -lambda.ind[i] + y[i]*log(lambda.ind[i]) - loggam(y[i]+1) )
    log(lambda.ind[i]) <- b1 + beta1*x1[i] + beta2*x2[i]
    logit(p[i]) <- b0 + gamma1*x1[i] + gamma2*x2[i]
  }
  lambda<- exp(beta1 + beta2)
  b0~dunif(-2.5,2.5)
  beta1~dnorm(0,tau1)
  gamma1~dnorm(0,tau2)
  beta2~dnorm(0,tau2)
  gamma2~dnorm(0,tau1)
  sigma1 <- b*(H1/(1-H1))
  sigma2 <- b*(H2/(1-H2))
```

```
  b <- 625
  H1~dbeta(0.5,0.5)
  H2~dbeta(0.5,0.5)
  tau1 <- 1/sigma1
  tau2 <- 1/sigma2
}

### Datos y valores iniciales ###
# y es el número de raíces
# x1 es la condición de fotoperiodo
# x2 es la concentración de la proteína BAP
# N es el número de datos
N <- 270
zip.data <- list("y","x1","x2","N")
zip.params <- c("beta1","gamma1","beta2","gamma2")
#Valores iniciales
zip.inits <- function(){list("beta1"=c(0.01),"beta2"=c(0.01),"gamma1"=c(0.01),
"gamma2"=c(0.01),"tau1"=c(10),"tau2"=c(10))}
# Generación de cadenas posteriores
zipfit <- bugs(data=zip.data, inits=zip.inits, parameters=zip.params, DIC=TRUE,
model.file = zip.model, n.chains = 1, n.iter = 10000, n.burnin = 2000, n.thin = 1)
attach.bugs(zipfit,overwrite = TRUE)
t=rbind(beta1,gamma1,beta2,gamma2,deviance)
f <- list(a=t,dic=zipfit$DIC)
detach.bugs()
g <- unlist(f,use.names = FALSE)
(dic <- g[40001])
l <- g[-40001]
beta1=l[seq(1,40000,5)]
gamma1=l[seq(2,40000,5)]
beta2=l[seq(3,40000,5)]
gamma2=l[seq(4,40000,5)]
deviance=l[seq(5,40000,5)]

### Análisis - Ajuste de una observación atípica ####
## A la variable respuesta se le adiciona un outlier
y <- c(rep(0,64),rep(1,10),rep(2,13),rep(3,15),rep(4,21),rep(5,18),rep(6,24),rep(7,21),
rep(8,23),rep(9,21),rep(10,17),rep(11,12),rep(12,5),rep(13,2),rep(14,3),34)
## Simulación de cadenas posteriores cuando los datos están contaminados con un outlier
zip.data <- list("y","x1","x2","N")
zip.params <- c("beta1","gamma1","beta2","gamma2")
#Valores iniciales
```

```
zip.inits <- function(){list("beta1"=c(0.01),"beta2"=c(0.01),"gamma1"=c(0.01),
"gamma2"=c(0.01),"tau1"=c(10),"tau2"=c(10))}
zipfit <- bugs(data=zip.data, inits=zip.inits, parameters=zip.params, DIC=TRUE,
model.file = zip.model, n.chains = 1, n.iter = 10000, n.burnin = 2000, n.thin = 1)
attach.bugs(zipfit,overwrite = TRUE)
t=rbind(beta1,gamma1,beta2,gamma2)
f <- list(a=t,dic=zipfit$DIC)
detach.bugs()
g <- unlist(f,use.names = FALSE)
l <- g[-32001]
beta1=l[seq(1,32000,4)]
gamma1=l[seq(2,32000,4)]
beta2=l[seq(3,32000,4)]
gamma2=l[seq(4,32000,4)]
## Mediana de los parámetros posteriores
b1 <- median(beta1)
b2 <- median(beta2)
g1 <- median(gamma1)
g2 <- median(gamma2)
## Lambda y p posteriores - A partir de la estructura del modelo ZIP
lambda <- exp(b1*8 + b2*2.2) # Para el outlier x1=8 y x2=2.2
library(boot)
p <- inv.logit(g1*8 + g2*2.2)
## Ajuste del outlier
(E_Y <- (1-p)*lambda)
## Diferencia absoluta entre el ajuste y el verdadero valor
abs(E_Y - 34)
```

Bibliografía

Agarwal, D., Gelfand, A. & Citron-Pousty, S. (2002), 'Zero-inflated models with application to spatial count data', *Environmental and Ecological Statistics* **9**, 341–355.

Angers, F. & Biswas, A. (2003), 'A Bayesian analysis of zero-inflated generalized Poisson model', *Computational Statistics and Data Analysis* **42**, 37–46.

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.

Cheung, Y. (2002), 'Zero-inflated models for regression analysis of count data: A study of growth and development', *Statistics in Medicine* **21**, 1461–1469.

Daniels, M. (1999), 'A Prior for the variance in hierarchical models', *The Canadian Journal of Statistics* **27**(3), 567–578.

Famoye, F. & Singh, K. (2006), 'Zero-inflated generalized Poisson regression model with an application to domestic violence data', *Journal of Data Science* **4**(1), 17–130.

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.

Gustafson, P., Hossain, S. & MacNab, Y. (2006), 'Conservative prior distributions for variance parameters in hierarchical models', *The Canadian Journal of Statistics* **34**(3), 377–390.

Hardin, J. & Hilbe, J. (2007), *Generalized Linear Models and Extensions*, segunda edn, Stata Press, College Station, Texas.

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

Karlis, D. & Ntzoufras, I. (2003), 'Analysis of sports data using bivariate Poisson models', *Journal of the Royal Statistical Society D* **52**, 381–393.

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. (2015), '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.

Pericchi, L. & Pérez, M. (2009), 'The case for a fully robust hierarchical Bayesian analysis of clinical trials.', *Technical report, Centro de Bioestadística y Bioinformática, Univ. Puerto Rico, Rio Piedras Campus. Disponible en: http://repositorio.upr.edu:8080/jspui/handle/10586%20/311* .

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

Rodrigues, J. (2003), 'Bayesian analysis of zero-inflated distributions', *Communications in Statistics - Theory and Methods* **32**(2), 281–289.

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

Xie, F., Lin, J. & Wei, B. (2014), 'Bayesian zero-inflated generalized Poisson regression model: estimation and case influence diagnostics', *Journal of Applied Statistics* **41**(6), 1383–1392.