Revista Colombiana de Estadística
Volumen 28 N° 1. pp. 17 a 21. Junio 2005

Una aproximación bayesiana al problema de heteroscedasticidad en el modelo lineal simple

Juan Carlos Correa*

Resumen

Presentamos una implementación bayesiana para ayudar a resolver un problema de heteroscedaticidad en el modelo de regresión simple, fácilmente extendible al caso múltiple.

***Palabras Claves**: Heteroscedasticidad, modelos de regresión, estadística bayesiana, muestreador de Gibbs.*

Abstract

We implement a bayesian solution to the heteroscedasticity problem in simple regression. This solution can be easily generalized to the multiple regression case.

***Keywords**: Heteroscedasticity, Regression models, Bayesian statistics, Gibbs sampler.*

1. Introducción

El modelo de regresión lineal es tal vez la herramienta estadística de más amplio uso. El modelo de regresión simple presenta una estructura elegante, útil para representar gran cantidad de fenómenos reales de una forma aproximada.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Se asume usualmente que $\epsilon_i \sim N(0, \sigma^2)$.

Un problema conocido es el de la heteroscedasticidad, o sea la violación de la condición de varianza constante de los errores.

Un modelo de heteroscedasticidad que se asume con frecuencia es $\sigma_i^2 = x_i^\nu \sigma^2$, donde ν se convierte en otro parámetro del modelo, el cual desde el punto de vista

*Profesor asociado. Escuela de Estadística. Universidad Nacional. Sede Medellín.
E-mail: jccorrea@unalmed.edu.co

tradicional se estima primero y se realiza una transformación antes de aplicar los procedimientos de estimación corrientes.

Aquí presentamos la aproximación bayesiana para la solución de este problema, mostrando cómo se puede generalizar al caso de la regresión múltiple y además lo ilustramos mediante un ejemplo utilizando el programa WinBUGS.

2. La aproximación bayesiana

Opuesto a la estimación clásica de parámetros, la estadística bayesiana produce distribuciones posteriores de las cantidades desconocidas (parámetros) teniendo en cuenta, tanto los datos, como las densidades *a priori* sobre estos parámetros. Como tal, la estadística bayesiana proporciona un cuadro más completo sobre la incertidumbre en la estimación de los parámetros desconocidos. Una introducción completa puede buscarse en el libro de Lee (1997). Si $\boldsymbol{\theta}$ es nuestro parámetro de interés y $\pi(\theta)$ es la distribución *a priori* de los parámetros que resume nuestra incertidumbre sobre los mismos, entonces, una vez obtenidos los datos, el paradigma bayesiano se establece como:

$$\pi(\boldsymbol{\theta}|\text{ datos}) \propto \pi(\theta)\boldsymbol{\theta}|\text{ datos})$$

donde L es la función de verosimilitud y $\propto$ es el símbolo de proporcionalidad, el cual indica que la cantidad debe dividirse por una constante para que sea una densidad propia.

Si $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ es la información muestral, la función de verosimilitud será:

$$\begin{aligned} L\left(\boldsymbol{\beta}, \sigma^2, \nu|\text{Datos}\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} x_i^{\nu/2} \sigma} \exp\left(-\frac{1}{2}\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{x_i^{\nu}\sigma^2}\right) \\ &= \frac{1}{(2\pi)^{n/2}\left(\prod_{i=1}^{n} x_i\right)^{\nu/2}\sigma} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{x_i^{\nu}\sigma^2}\right) \end{aligned}$$

Si asumimos una distribución *a priori* no informativa sobre $\left(\boldsymbol{\beta}, \sigma^2, \nu\right)$ tal que:

$$\begin{aligned} \pi(\boldsymbol{\beta}) &\propto c \\ \pi(\sigma^2) &\propto \frac{1}{\sigma} \\ \pi(\nu) &\propto k \end{aligned}$$

donde c y k son constantes, entonces la distribución posterior de $(\boldsymbol{\beta}, \sigma^2, \nu)$ es:

$$\pi\left(\boldsymbol{\beta}, \sigma^2, \nu|Datos\right) \propto \frac{1}{\left(\prod_{i=1}^{n} x_i\right)^{\nu/2}\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{x_i^{\nu}}\right)$$

La anterior expresión puede resolverse utilizando métodos computacionales conocidos como MCMC (Monte Carlo Markov Chain) utilizados ahora para resolver problemas bayesianos de gran complejidad. Uno de los procedimientos es

el muestreador de Gibbs, técnica introducida por Geman & Geman (1984) y desarrolladas posteriormente por Gelfand & Smith (1990). En términos generales, la idea del muestreador de Gibbs es la de sobreponer las dificultades del cálculo de $\pi\left(\boldsymbol{\beta}, \sigma^2, \nu|\text{Datos}\right)$ con una sucesión de cálculos más fáciles de distribuciones condicionales que funciona así:

1. Seleccionar un punto arbitrario $\left(\boldsymbol{\beta}_0, \sigma_0^2, \nu_0\right)$.
2. Generar:
   a) $\boldsymbol{\beta}_i$ de $\pi\left(\boldsymbol{\beta}|Datos, \sigma_{i-1}^2, \nu_{i-1}\right)$.
   b) σ_i^2 de $\pi\left(\sigma^2|Datos, \boldsymbol{\beta}_i, \nu_{i-1}\right)$.
   c) ν_i de $\pi\left(\nu|Datos, \boldsymbol{\beta}_i, \sigma_i^2\right)$.
3. Repetir el paso anterior un número grande veces.

Dado que este proceso de generación de muestras es un proceso markoviano donde la distribución estacionaria es la distribución posterior de la cual se pretende extraer las muestras, se deben descartar los valores generados al comienzo del proceso. Este es un problema complejo pero se acostumbra eliminar los primeros 1000 valores y generar 5000 valores o más.

Es fácil probar que en el paso i, la distribución $\pi\left(\boldsymbol{\beta}|\text{Datos}, \sigma_{i-1}^2, \nu_{i-1}\right)$ es una $N\left(\hat{\boldsymbol{\beta}}, \sigma_{i-1}^2\left(\boldsymbol{X^*}'\boldsymbol{X^*}\right)^{-1}\right)$, donde la j-ésima fila de $\boldsymbol{X^*}$ es el vector

$$\left(\frac{1}{\sigma_{i-1}x_j^{\nu_{(i-1)}/2}}, \frac{x_j}{\sigma_{i-1}x_j^{\nu_{(i-1)}/2}}\right)$$

y $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X^*}'\boldsymbol{X^*}\right)^{-1}\boldsymbol{X^*}'\boldsymbol{y^*}$, donde la j-ésima componente de $\boldsymbol{y^*}$ es

$$y_j^* = \frac{y_j}{\left(\sigma_{i-1}x_j^{\frac{\nu_{(i-1)}}{2}}\right)}.$$

La distribución posterior condicional de σ^2 es una gamma inversa:

$$\pi(\sigma^2 \mid \text{Datos}, \boldsymbol{\beta}_i, \nu_{i-1}) = IG\left(2, \quad \frac{1}{2}\sum_{j=1}^{n}\frac{(y_j - \beta_{0_i} - \beta_{1_i}x_j)^2}{x_j^{\nu_{i-1}}}\right)$$

La distribución posterior condicional de ν no tiene una forma cerrada y por lo tanto debemos utilizar algún método para la generación de valores de esta distribución, tal como el muestreo por importancia u otro, ya que:

$$\pi(\nu \mid \text{Datos}, \boldsymbol{\beta}_i, \sigma_i^2) = \frac{1}{(\prod_{i=1}^{n}x_i)^{\nu}}\exp\left(-\frac{1}{2\sigma_i^2}\sum_{i=1}^{n}\frac{(y_i - \beta_{0_i} - \beta_{1_i}x_i)^2}{x_i^{\nu}}\right)$$

3. Ejemplo

La modelación del precio del metro cuadrado de apartamentos usados es un problema de interés para las lonjas de propiedad raíz. Consideramos el modelo:

$$\text{Precio} = \beta_0 + \beta_1 Mts2$$

que relaciona el precio de oferta de un apartamento usado y su tamaño en metros cuadrados. Dada la complejidad del problema sólo consideramos apartamentos de un sector limitado de la ciudad de Medellín como es *El Poblado*, donde tienen una estructura homogénea: estratos socioeconómicos altos, calidad de construcciones similares, etc. Realizamos el ajuste del modelo con el programa WinBUGS, versión 1.4. Como este programa no permite trabajar con distribuciones *a priori* no informativas, seleccionamos distribuciones con varianzas muy grandes con el fin de imitar dentro de un rango plausible a las no informativas. El código aparece a continuación:

```
model
{
  for( i in 1:N ) {
    Precio[i] ~ dnorm(mu[i],precision[i])
    mu[i] <- alpha.star + beta*(Mts2[i] - mean(Mts2[]))
    precision[i] <- tauC/exp(log(Mts2[i])*nu)
  }
  alpha <- alpha.star - beta * mean(Mts2[])
  beta ~ dnorm(0.0,0.001)
  alpha.star ~ dnorm(0.0,0.001)
  nu ~ dnorm(0.0,0.001)
  tauC ~ dgamma(0.001, 0.001)
  desv<-1/sqrt(tauC)
}
```

```
list(Mts2=c( 91 ,228 ,152 ,130 ,130 ,127 ,165 ,213 ,250 ,
240 ,200 ,260 ,85 ,300 ,76 ,97 ,236 ,315 ,118.5 ,140 ,
225 ,225 ,207 ,193 ,158 ,80 ,114 ,190 ,99.9 ,140 ,75 ,
210 ,145 ,148 ,86 ,88 ,156 ,110 ,140 ,99 ,175 ,435 ,179 ,
100 ,79 ,100 ,167 ,88 ,223 ,218 ,200 ,130 ,75 ,169.2 ,
170 ,158.2 ),
Precio=c( 132 ,175 ,106.4 ,120 ,140 ,134 ,180 ,200 ,235 ,240 ,
260 ,240 ,75 ,190 ,86 ,93 ,242 ,295 ,102 ,180 ,310 ,280 ,
224 ,155 ,160 ,100 ,138 ,133 ,110 ,153 ,60 ,280 ,140 ,215 ,
75 ,83 ,125 ,80 ,135 ,145 ,190 ,410 ,150 ,115 ,97 ,100 ,
175 ,80 ,395 ,235 ,230 ,137 ,125 ,270 ,300 ,140 ),
N=56)
```

```
list(alpha.star=0, beta=0,nu=0,tauC=1)
```

Se generaron 13000 muestras de las cuales se descartaron las primeras 4000 y para el proceso inferencial se utilizaron 9000. La tabla presenta algunos resúmenes de las distribuciones posteriores de los cuatro parámetros del modelo. Se observa como el intervalo de probabilidad para ν del 95 % es (1,22, 2,59), lo cual muestra cómo es de grave el problema de la varianza no constante en el modelo.

	media	sd	error MC	2.5 %	mediana	97.5 %
ν	1,848	0,3767	0,03346	1,22	1,807	2,592
α	12,79	12,52	0,1125	−11,7	12,68	37,7
β	0,9473	0,0957	$9{,}03E-4$	0,7549	0,949	1,13
σ	0,5901	0,5272	0,0437	0,06027	0,4378	1,972

4. Conclusiones y recomendaciones

La aproximación bayesiana permite resolver el problema de varianza no constante en regresión en un solo paso, ya que simultáneamente se hallan las distribuciones posteriores de todos los parámetros involucrados, un problema que para la aproximación tradicional es complejo y requiere primero modelar la heteroscedasticidad y luego estimar los parámetros del modelo realizando los ajustes necesarios.

Esta solución se extiende fácilmente al caso de la regresión múltiple, ya que simplemente una estructura como $\sigma_i^2 = x_{i1}^{\nu_1} x_{i2}^{\nu_2} \cdots x_{ik}^{\nu_k} \sigma^2$, se modela en forma similar.

Bibliografía

Gelfand, A. & Smith, A. F. M. (1990), 'Sampling based approaches to calculating marginal densities', *J. Amer. Stat. Assoc* **85**, 398–409.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions and the bayesian restoration of images', *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Lee, P. M. (1997), *Bayesian Statistics: An Introduction*, 2 edn, Arnold, London.

Tanner, M. A. (1996), *Tools for Statistical Inference*, Springer-Verlag, New York.