Negative Binomial Generalized Linear Model for Zero-truncated Count Data

Brian M. Brost

23 March 2016

Description

A generalized linear model for zero-truncated count data.

Implementation

The file zt.nb.glm.sim.R simulates data according to the model statement presented below, and zt.nb.glm.mcmc.R contains the MCMC algorithm for model fitting.

Derivation of zero-truncated negative binomial distribution

The probability mass function of the (non-truncated) negative binomial distribution is:

$$[z] = \left(\frac{\Gamma(z+\alpha)}{\Gamma(a)\,\Gamma(z+1)}\right)\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(1-\frac{\alpha}{\lambda+\alpha}\right)^{z} \tag{1}$$

It follows that the probability that $z=0$ is

$$\begin{aligned}
[z \mid z=0] &= \left(\frac{\Gamma(\alpha)}{\Gamma(a)\,\Gamma(1)}\right)\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(1-\frac{\alpha}{\lambda+\alpha}\right)^{0} && (2)\\
&= \left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}, && (3)
\end{aligned}$$

and thus the probability that $z>0$ is $1-[z \mid z=0] = 1-\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}$. We arrive at the density function of the zero-truncated negative binomial distribution by excluding the probability that $z=0$ from the standard negative binomial distribution (Eq. 1). This is accomplished by dividing Eq. 1 by $[z \mid z=0]$:

$$\begin{aligned}
[z \mid z>0] &= \left(\frac{\Gamma(z+\alpha)}{\Gamma(a)\,\Gamma(z+1)}\right)\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(1-\frac{\alpha}{\lambda+\alpha}\right)^{z}\left(1-\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\right)^{-1} && (4)\\
&= \mathrm{NB}(z \mid \lambda,\alpha)\left(1-\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\right)^{-1}. && (5)
\end{aligned}$$

We abbreviate the density function for the zero-truncated negative binomial distribution as $\mathrm{ZTNB}(\lambda,\alpha)$.

Model statement

Let z_i, for $i=1,\ldots,n$, be observed non-zero count data (i.e., z_i are integers greater than 0). Also let $\mathbf{x}_i$ be a vector of covariates associated with z_i for which inference is desired, and the vector $\boldsymbol{\beta}$ be the corresponding coefficients.

$$\begin{aligned}
z_i &\sim \mathrm{ZTNB}(\lambda_i,\alpha)\\
\log(\lambda_i) &= \mathbf{x}_i'\boldsymbol{\beta}\\
\boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0},\sigma_\beta^2\mathbf{I})\\
\alpha &\sim \mathrm{Gamma}(a,b),
\end{aligned}$$

where $\mathrm{E}[z_i]=\lambda_i$ and $\mathrm{Var}[z_i]=\lambda_i+\frac{\lambda_i^2}{\alpha}$.

Full conditional distributions

Regression coefficients ($\boldsymbol{\beta}$):

$$
\begin{aligned}
[\boldsymbol{\beta} \mid \cdot] &\propto \prod_{i=1}^{n} [z_i \mid \boldsymbol{\beta}, \alpha] \, [\boldsymbol{\beta}] \\
&\propto \prod_{i=1}^{n} \mathrm{ZTNB} \left(z_i \mid \mathbf{x}_i' \boldsymbol{\beta}, \alpha \right) \mathcal{N} \left(\boldsymbol{\beta} \mid \mathbf{0}, \sigma_\beta^2 \mathbf{I} \right).
\end{aligned}
$$

The update for $\boldsymbol{\beta}$ proceeds using Metropolis-Hastings.

Dispersion (i.e., size) parameter (α):

$$
\begin{aligned}
[\alpha \mid \cdot] &\propto \prod_{i=1}^{n} [z_i | \boldsymbol{\beta}, \alpha] \, [\alpha] \\
&\propto \prod_{i=1}^{n} \mathrm{ZTNB} \left(z_i \mid \mathbf{x}_i' \boldsymbol{\beta}, \alpha \right) \mathrm{Gamma} \left(\alpha \mid a, b \right).
\end{aligned}
$$

The update for α proceeds using Metropolis-Hastings.