Notes on Hurdle Negative Binomial Regression

Adam Kapelner[1]

[1]*Department of Mathematics, Queens College, CUNY, USA*

August 21, 2023

We begin with examining the zero-inflated model and then talk about why it's wiser to go with the hurdle model.

So we start by modeling count data using a negative binomial model where the zeroes are inflated. Let inflation be defined as the latent variable $I_i = 1$ and uninflated be $I_i = 0$. We first define the probability p_i of an inflated zero using a generalized linear model (GLM) as

$$p_i := \mathbb{P}\left(Y_i = 0 \mid I_i = 1\right) := \left(1 + \exp\left(-\eta_i\right)\right)^{-1}$$

where

$$\eta_i := \gamma_0 + \gamma_1 x_{i,1} + \ldots + \gamma_p x_{i,p}$$

and conveniently

$$1 - p_i = \left(1 + \exp\left(\eta_i\right)\right)^{-1}.$$

We then define the count model for y_i which is uninflated $I_i = 0$ as the negative binomial model parameterized with the mean as stated here and a generalized linear model (GLM):

$$\begin{aligned}
\mathbb{P}\left(Y_i = y_i \mid I_i = 0\right) &= \binom{y_i + \phi - 1}{y_i}\left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i}\left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \\
&= \frac{\Gamma\left(y_i + \phi - 2\right)}{y_i!\,\Gamma\left(\phi - 2\right)}\left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i}\left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \\
&= \frac{\Gamma\left(y_i + \phi - 2\right)}{y_i!\,\Gamma\left(\phi - 2\right)}\left(\frac{\exp\left(\xi_i\right)}{\exp\left(\xi_i\right) + \phi}\right)^{y_i}\left(\frac{\phi}{\exp\left(\xi_i\right) + \phi}\right)^{\phi} \\
&= \frac{\Gamma\left(y_i + \phi - 2\right)}{y_i!\,\Gamma\left(\phi - 2\right)}\left(1 + \phi \exp\left(-\xi_i\right)\right)^{-y_i}\left(1 + \phi^{-1}\exp\left(\xi_i\right)\right)^{-\phi}
\end{aligned}$$

where

$$\xi_i := \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}.$$

Which means the probability of any realization would be

$$\mathbb{P}\left(Y_i = y_i\right) = \mathbb{P}\left(Y_i = 0 \mid I_i = 1\right) \mathbb{1}_{y_i=0} + (1 - \mathbb{P}\left(Y_i = 0 \mid I_i = 1\right))\mathbb{P}\left(Y_i = y_i \mid I_i = 0\right).$$

The problem is that plus sign will destroy the optimization because you can't log it effectively. Let's now consider the hurdle model.

Here, there is a probability of zero. And if it "jumps the hurdle" then we get a positive realization model. We can still model the positive realizations with a negative binomial model by just subtracting one from the counts to shift the support from $\{1, 2, \ldots\}$ to $\{0, 1, \ldots\}$. The hurdle is then defined as before where this time there is no latent "inflation" variable:

$$\begin{aligned} p_i &:= \mathbb{P}\left(Y_i = 0\right) := \left(1 + \exp\left(-\eta_i\right)\right)^{-1} \\ 1 - p_i &= \left(1 + \exp\left(\eta_i\right)\right)^{-1} \end{aligned}$$

The positive realization model is then defined as before except now we subtract one from every y_i to shift the support correctly. FROM THIS POINT ON, all $y_i \geq 1$.

$$\begin{aligned} \mathbb{P}\left(Y_i = y_i \mid Y_i > 0\right) &= \binom{y_i + \phi - 2}{y_i - 1} \left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i - 1} \left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \\ &= \frac{\Gamma\left(y_i + \phi - 3\right)}{(y_i - 1)!\,\Gamma\left(\phi - 2\right)} \left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i - 1} \left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \\ &= \frac{\Gamma\left(y_i + \phi - 3\right)}{(y_i - 1)!\,\Gamma\left(\phi - 2\right)} \left(\frac{\exp\left(\xi_i\right)}{\exp\left(\xi_i\right) + \phi}\right)^{y_i - 1} \left(\frac{\phi}{\exp\left(\xi_i\right) + \phi}\right)^{\phi} \end{aligned}$$

Which means the probability of any realization would be

$$\mathbb{P}\left(Y_i = y_i\right) = \mathbb{P}\left(Y_i = 0\right) \mathbb{1}_{y_i=0} + (1 - \mathbb{P}\left(Y_i = 0\right))\mathbb{P}\left(Y_i = y_i \mid Y_i > 0\right) \mathbb{1}_{y_i>0}.$$

We can make life easier by defining the augmented data $z_i := \mathbb{1}_{y_i=0}$ to obtain:

$$\begin{aligned} \mathbb{P}\left(Y_i = y_i, Z_i = z_i\right) &= \mathbb{P}\left(Y_i = 0\right)^{z_i} \left((1 - \mathbb{P}\left(Y_i = 0\right))\mathbb{P}\left(Y_i = y_i \mid Y_i > 0\right)\right)^{1 - z_i} \\ &= p_i^{z_i} \left((1 - p_i)\mathbb{P}\left(Y_i = y_i \mid Y_i > 0\right)\right)^{1 - z_i} \end{aligned}$$

The total likelihood function will be:

$$\begin{aligned} & \mathcal{L}(\gamma_0, \gamma_1, \ldots, \gamma_p, \beta_0, \beta_1, \ldots, \beta_p, \phi \,|\, y_1, \ldots, y_n, z_1, \ldots, z_n) \\ = & \prod_{i=1}^{n} p_i^{z_i} \left((1 - p_i)\mathbb{P}\left(Y_i = y_i \mid Y_i > 0\right)\right)^{1 - z_i} \end{aligned}$$

And the log-likelihood will be

$$\ell(\gamma_0, \gamma_1, \ldots, \gamma_p, \beta_0, \beta_1, \ldots, \beta_p, \phi \,|\, y_1, \ldots, y_n, z_1, \ldots, z_n)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} z_i \ln(p_i) + (1-z_i)\ln(1-p_i) + (1-z_i)\ln\left(\mathbb{P}(Y_i = y_i \mid Y_i > 0)\right) \\
&= \sum_{i=1}^{n} z_i \ln\left((1+\exp(-\eta_i))^{-1}\right) + (1-z_i)\ln\left((1+\exp(\eta_i))^{-1}\right) + \\
&\quad (1-z_i)\ln\left(\frac{\Gamma(y_i+\phi-3)}{(y_i-1)!\,\Gamma(\phi-2)}(1+\phi\exp(-\xi_i))^{-(y_i-1)}\left(1+\phi^{-1}\exp(\xi_i)\right)^{-\phi}\right) \\
&= \sum_{i=1}^{n} -z_i \ln(1+\exp(-\eta_i)) - (1-z_i)\ln(1+\exp(\eta_i)) + \\
&\quad (1-z_i)\ln\left(\frac{\Gamma(y_i+\phi-3)}{(y_i-1)!\,\Gamma(\phi-2)}\right) + \\
&\quad (1-z_i)\ln\left((1+\phi\exp(-\xi_i))^{-(y_i-1)}\right) + \\
&\quad (1-z_i)\ln\left(\left(1+\phi^{-1}\exp(\xi_i)\right)^{-\phi}\right) \\
&= \sum_{i=1}^{n} -z_i \ln(1+\exp(-\eta_i)) + \\
&\quad -(1-z_i)\ln(1+\exp(\eta_i)) + \\
&\quad (1-z_i)\left(\ln\Gamma(y_i+\phi-3) - \ln\Gamma(y_i) - \ln\Gamma(\phi-2)\right) + \\
&\quad -(y_i-1)(1-z_i)\ln(1+\phi\exp(-\xi_i)) + \\
&\quad -\phi(1-z_i)\ln\left(1+\phi^{-1}\exp(\xi_i)\right)
\end{aligned}
$$

To simplify this a little bit, note that when $z_i = 1$, the summand simplifies to:

$$
-\ln(1+\exp(-\eta_i))
$$

And when $z_i = 0$, the summand simplifies to:

$$
\begin{aligned}
&-\ln(1+\exp(\eta_i)) + \\
&\ln\Gamma(y_i+\phi-3) - \ln\Gamma(y_i) - \ln\Gamma(\phi-2) + \\
&-(y_i-1)\ln(1+\phi\exp(-\xi_i)) + \\
&-\phi\ln\left(1+\phi^{-1}\exp(\xi_i)\right)
\end{aligned}
$$

where once again the approximation follows from a Taylor series approximation.

We seek to maximize this quantity over the parameters. We can start the parameters from an intelligent point by fitting a logistic regression to the z_i's and returning a starting point for the γ_j's. Then we can fit an OLS model to the y_i's which are nonzero returning a starting point for the β_j's.

When using the L-BFGS algorithm, we also need the gradient $\nabla\ell$ with respect to all of our parameters, i.e. $\gamma_0, \gamma_1, \ldots, \gamma_p, \beta_0, \beta_1, \ldots, \beta_p, \phi$. We now derive them. Assume the first column of the covariate matrix is 1. First when $z_i = 1$,

$$\frac{\partial \ell}{\partial \gamma_k} \quad := \quad x_{i,k}\left(1+\exp\left(\eta_i\right)\right)^{-1}$$

and all other gradients are zero. Then when $z_i = 0$,

$$\begin{aligned}
\frac{\partial \ell}{\partial \gamma_k} \quad &:= \quad -x_{i,k}\left(1+\exp\left(-\eta_i\right)\right)^{-1} + \\
&\quad\;\; -\phi\left(1+\phi\exp\left(-\xi_i\right)\right)^{-1} \\
\frac{\partial \ell}{\partial \beta_k} \quad &:= \quad -x_{i,k}(y_i-1)\phi\left(\phi+\exp\left(\xi_i\right)\right)^{-1} + \\
&\quad\;\; -x_{i,k}\phi\left(1+\phi\exp\left(-\xi_i\right)\right)^{-1} \\
\frac{\partial \ell}{\partial \phi} \quad &:= \quad \psi(y_i+\phi-3)-\psi(\phi-2) + \\
&\quad\;\; -(y_i-1)\left(\phi+\exp\left(\xi_i\right)\right)^{-1} + \\
&\quad\;\; -\ln\left(\phi+\exp\left(\xi_i\right)\right)+\left(1+\phi\exp\left(-\xi_i\right)\right)^{-1}+\ln\left(\phi\right)
\end{aligned}$$

where ψ denots the digamma function.

One of the problems is that many of the terms above can suffer from numerical over/underflow due to the exponentiation. Here are some Taylor series approximations for such computations up the fifth order:

$$\begin{aligned}
\ln\left(1+\exp\left(x\right)\right) \quad &\approx \quad \ln\left(2\right)+\frac{x}{2}+\frac{x^2}{8}-\frac{x^4}{192} \\
\ln\left(1+\exp\left(-x\right)\right) \quad &\approx \quad \ln\left(2\right)-\frac{x}{2}+\frac{x^2}{8}-\frac{x^4}{192} \\
\ln\left(1+c\exp\left(x\right)\right) \quad &= \quad \ln\left(1+\exp\left(\ln\left(c\right)+x\right)\right) \\
\ln\left(1+c\exp\left(-x\right)\right) \quad &= \quad \ln\left(1+\exp\left(\ln\left(c\right)-x\right)\right) \\
\ln\left(c+\exp\left(x\right)\right) \quad &= \quad \ln\left(c\right)+\ln\left(1+\exp\left(x-\ln\left(c\right)\right)\right) \\
\left(1+\exp\left(x\right)\right)^{-1} \quad &\approx \quad \frac{1}{2}-\frac{x}{4}+\frac{x^3}{48}-\frac{x^5}{480} \\
\left(1+\exp\left(-x\right)\right)^{-1} \quad &\approx \quad \frac{1}{2}+\frac{x}{4}-\frac{x^3}{48}+\frac{x^5}{480} \\
\left(1+c\exp\left(x\right)\right)^{-1} \quad &= \quad \left(1+\exp\left(\ln\left(c\right)+x\right)\right)^{-1} \\
\left(1+c\exp\left(-x\right)\right)^{-1} \quad &= \quad \left(1+\exp\left(\ln\left(c\right)-x\right)\right)^{-1} \\
\left(c+\exp\left(x\right)\right)^{-1} \quad &= \quad \frac{1}{c}\left(1+c^{-1}\exp\left(x\right)\right)^{-1}
\end{aligned}$$