

# Notes on Hurdle Negative Binomial Regression

ADAM KAPELNER<sup>1</sup>

<sup>1</sup>*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

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

where

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

and conveniently

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

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):

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

where

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

Which means the probability of any realization would be

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

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  $\{1, 2, \dots\}$  to  $\{0, 1, \dots\}$ . The hurdle is then defined as before where this time there is no latent "inflation" variable:

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

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  $y_i \geq 1$ .

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

Which means the probability of any realization would be

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

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

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

The total likelihood function will be:

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

And the log-likelihood will be

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

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

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:

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

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  $\gamma_j$ 's. Then we can fit an OLS model to the  $y_i$ 's which are nonzero returning a starting point for the  $\beta_j$ 's.

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

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

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

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

where  $\psi$  denotes 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:

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