Hurdle negative binomial regression model with right censored count data

Seyed Ehsan Saffari[1], Robiah Adnan[2] and William Greene[3]

Abstract

A Poisson model typically is assumed for count data. In many cases because of many zeros in the response variable, the mean is not equal to the variance value of the dependent variable. Therefore, the Poisson model is no longer suitable for this kind of data. Thus, we suggest using a hurdle negative binomial regression model to overcome the problem of overdispersion. Furthermore, the response variable in such cases is censored for some values. In this paper, a censored hurdle negative binomial regression model is introduced on count data with many zeros. The estimation of regression parameters using maximum likelihood is discussed and the goodness-of-fit for the regression model is examined.

1. Introduction

Commonly, for the modelling of counts such as the number of reported insurance claims, the starting point is the Poisson distribution:

$$f_{Y_i}(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \tag{1}$$

where covariates are included in the model by the parameter $\lambda_i = \exp(x_i^\mathsf{T} \beta)$ where x_i is a vector of explanatory variables (Dionne and Vanasse, 1989). The Poisson distribution

---

[1,2] Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.

[3] Department of Economics, Stern School of Business, New York University 44 West 4th St., New York, NY, 10012, USA.

is equidispersed since its mean and variance are both equal to λ_i. Because the Poisson distribution has some severe drawbacks that limit its use, other distributions can be used, such as hurdle models (Boucher *et al.*, 2007).

Mullahy (1986) has first discussed hurdle count data models. Hurdle models allow for a systematic difference in the statistical process governing individuals (observations) below the hurdle and individuals above the hurdle. In particular, a hurdle model is mixed by a binary outcome of the count being below or above the hurdle (the selection variable), with a truncated model for outcomes above the hurdle. That is why hurdle models sometimes are also called two-part models.

The most important usage of a hurdle count data model is the hurdle at zero. The hurdle at zero formulation can account for excess zeros. It means that this model can be used in situations where there are many zeros at the response variable. In this case, the hurdle at zero defines a probability $(\Pr(Y = 0))$ that is the first part of the two part-models.

The hurdle model is flexible and can handle both under- and overdispersion problem. A generalized hurdle model is introduced by Gurmu (1998) for the analysis of overdispersed or underdispersed count data. Greene (2005) has discussed about the comparison between hurdle and zero-inflated models as two part-models. Some researchers have discussed the applications of hurdle models, such as Pohlmeier and Ulrich (1995), Arulampalam and Booth (1997). A hurdle model to the annual number of recreational boating trips by a family is discussed by Gurmu and Trivedi (1996). Dalrymple, Hudson and Ford (2003) applied three mixture models including a hurdle model and argued its application in the incidence of sudden infant death syndrome (SIDS). Boucher, Denuit and Guillen (2007) compared generalized heterogeneous, zero-inflated, hurdle, and compound frequency models for the annual number of claims reported to the insurer. Saffari, Adnan and Greene (2011) argued the overdispersion problem on count data using a right truncated Poisson regression model.

Suppose that $g_1(0)$ is the probability value when the value for response variable is zero and that $g_2(k), k = 1, 2, \ldots$ is a probability function when the response variable is a positive integer. Therefore, the probability function of the hurdle-at-zero model is given by:

$$p(Y_i = k) = \begin{cases} g_1(0), & k = 0, \\ (1 - g_1(0)) \, g_2(k), & k = 1, 2, \ldots \end{cases} \tag{2}$$

Mullahy (1986) discussed the hurdle-at-zero model and he believes that both parts of the hurdle model are based on probability functions for nonnegative integers such as f_1 and f_2. In terms of the general model above, let $g_1(0) = f_1(0)$ and $g_2(k) = f_2(k)/(1 - f_2(0))$. In the case of g_2, normalization is required because f_2 has support over the nonnegative integers ($k = 0, 1, \ldots$) whereas the support of g_2 must be over the positive integers ($k = 1, 2, \ldots$). This means that we need to truncate the probability function f_2. However, this is a theoretical concept, i.e., truncation on f_2 does not mean

that there is truncation of the population here. All we need to do is to work with a distribution with positive support, and the second part of a hurdle model can use a displaced distribution or any distribution with positive support as well.

Under the Mullahy (1986) assumptions, the probability distribution of the hurdle-at-zero model is given by

$$f(Y = 0) = f_1(0)$$

$$f(Y = k) = \frac{1 - f_1(0)}{1 - f_2(0)} f_2(k) = \theta f_2(k), \quad k = 1, 2, \ldots$$

where f_2 is referred to as *parent*-process. The numerator of θ presents the probability of crossing the hurdle and the denominator gives a normalization that accounts for the (purely technical) truncation of f_2. It follows that if $f_1 = f_2$ or, equivalently, $\theta = 1$ then the hurdle model collapses to the parent model. The expected value of the hurdle model is given by

$$E(Y) = \theta \sum_{k=1}^{\infty} k f_2(k) \tag{3}$$

and the difference between this expected value and the expected value of the parent model is the factor θ. In addition, the variance value of the hurdle model is given by

$$Var(Y) = \theta \sum_{k=1}^{\infty} k^2 f_2(k) - \left[\theta \sum_{k=1}^{\infty} k f_2(k) \right]^2 \tag{4}$$

If θ exceeds 1, it means that the probability of crossing the hurdle is greater than the sum of the probabilities of positive outcomes in the parent model. Therefore, increasing the expected value of the hurdle model is related to the expected value of the parent model. Alternatively, if θ is less than 1 (that is the usual case in an application with excess zeros), it means that the probability of not crossing the hurdle is greater than the probability of a zero in the parent model, thus decreasing the expected value of the hurdle model relatively to the expected value of the parent model. Therefore, this model gives a new explanation of excess zeros as being a characteristic of the mean function rather than a characteristic of the variance function. The mean function of the hurdle model provides additional nonlinearities relative to the standard model in order to account for the corner solution outcome, much as in other corner solution models, and this is just like as how a Tobit model works.

Consequently, the model can be overdispersed and that depends on the value of the parent processes. To overcome overdispersion, we would like to cut the values of the response variable that are very big. In statistics, this is called truncation and because we want to truncate the values that are bigger than a constant, it is called a right truncation.

There are many options to choose the processes f_1 and f_2. Some of the most popular hurdle model choices are nested models where f_1 and f_2 come from the same distribution, such as the Poisson distribution (Mullahy, 1986) or the Negative Binomial (Pohlmeier and Ulrich, 1995). However, non-nested models (Grootendorst (1995), (Gurmu, 1998), or Winkelmann (2003)) can also be used. These models do not nest with a standard count distributions such as the Poisson or the NB types, but are overlapping with suggested model by Vuong (1989) since models can be equivalent for certain parameter restrictions.

Maximum likelihood is used to estimate the parameters. The log-likelihood function of a hurdle model can be expressed as:

$$ll = \sum_{i=1}^{n} I_{\{y_i=0\}} \log f_1(0; \theta_i) + I_{\{y_i>0\}} \log(1 - f_1(0; \theta_i)) + \sum_{i=1}^{n} I_{\{y_i>0\}} \log \frac{f_2(y_i; \theta_i)}{1 - f_2(0; \theta_i)} \quad (5)$$

The interesting aspect of the hurdle model is to estimate the parameters by two separate steps. In fact, we can estimate the zero-part parameters by using MLE on the first part of the likelihood function while the other parameters only use the second part, only composed with non-zero elements. We have used SAS code to implement this algorithm and this characteristic of the model helps us to save computer time in the estimation (Chou and Steenhard, 2009).

In this article, the main objective is to explain how we can use hurdle negative binomial regression model in right censored data. In Section 2, the hurdle negative binomial regression model is defined and the likelihood function of hurdle negative binomial regression model in right censored data is formulated. In Section 3, the parameter estimation is discussed using maximum likelihood. In Section 4, the goodness-of-fit for the regression model is examined and a test statistic for examining the dispersion of regression model in right censored data is proposed. An example is conducted for a censored hurdle negative binomial regression model in terms of the parameter estimation, standard errors and goodness-of-fit statistic in Section 5.

2. The model

Let $Y_i (i = 1, 2, \ldots, n)$ be a nonnegative integer-valued random variable and suppose $Y_i = 0$ is observed with a frequency significantly higher than can be modeled by the usual model. We consider a hurdle negative binomial regression model in which the response variable $Y_i (i = 1, \ldots, n)$ has the distribution

$$\Pr(Y_i = y_i) = \begin{cases} w_0, & y_i = 0, \\ (1 - w_0) \dfrac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \dfrac{(1 + \alpha\mu_i)^{-\alpha^{-1} - y_i} \alpha^{y_i} \mu_i^{y_i}}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}}, & y_i > 0, \end{cases} \quad (6)$$

or

$$\Pr(Y_i = y_i) = \begin{cases} w_0, & y_i = 0, \\ (1 - w_0)\dfrac{g}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}}, & y_i > 0, \end{cases} \tag{7}$$

where

$$g = g(y_i; \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}(1 + \alpha\mu_i)^{-\alpha^{-1} - y_i}\alpha^{y_i}\mu_i^{y_i} \tag{8}$$

where $\alpha(\geq 0)$ is a dispersion parameter that is assumed not to depend on covariates. In addition, we suppose $0 < w_0 < 1$ and $w_0 = w_0(z_i)$ satisfy

$$\text{logit}(w_0) = \log(\frac{w_0}{1 - w_0}) = \sum_{j=1}^{m} z_{ij}\delta_j \tag{9}$$

where $(z_{i1} = 1, z_{i2}, \ldots, z_{im})$ is the i-th row of covariate matrix Z and $(\delta_1, \delta_2, \ldots, \delta_m)$ is an unknown m-dimensional column vector of parameters. In this set up, the non-negative function w_0 is modeled via logit link function. This function is linear and other appropriate link functions that allow w_0 being negative may be used. In addition, there is interest in capturing any systematic variation in μ_i, the value of μ_i is most commonly placed within a loglinear model

$$\log(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j \tag{10}$$

and β_j's are the independent variables in the regression model and m is the number of these independent variables. Furthermore, in this paper we suppose that w_0 and β_j are not related.

The value of response variable, Y_i, for some observations in a data set, may be censored. If censoring occurs for the ith observation, we have $Y_i \geq y_i$ (right censoring). However, if no censoring occurs, we know that $Y_i = y_i$. Thus, we can define an indicator variable d_i as

$$d_i = \begin{cases} 1 & \text{if } Y_i \geq y_i, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

We can now write

$$\Pr(Y_i \geq y_i) = \sum_{j=y_i}^{\infty} \Pr(Y_i = j) = 1 - \sum_{j=0}^{y_i - 1} \Pr(Y_i = j) \tag{12}$$

Therefore, the log-likelihood function of the censored hurdle regression model can be written as

$$\log L(\theta_i; y_i) = \sum_{i=1}^{n} \left\{ (1 - d_i) \left[I_{y_i=0} \log f(0; \theta_i) + I_{y_i>0} f(y_i; \theta_i) \right] + d_i \log \left(\sum_{j=y_i}^{\infty} \Pr(Y_i = j) \right) \right\}$$
(13)

We now obtain the log-likelihood function for the hurdle negative binomial regression model, we have

$$LL = \sum_{i=1}^{n} \left\{ (1 - d_i) \left[I_{y_i=0} \log w_0 + I_{y_i>0} \left\{ \log(1 - w_0) + \log g - \log \left(1 - (1 + \alpha \mu_i)^{-\alpha^{-1}} \right) \right\} \right] \right.$$
$$\left. + d_i \log \sum_{j=y_i}^{\infty} \Pr(Y_i = j) \right\}$$
(14)

3. Parameter estimation

In this section we estimate the parameters by maximum likelihood. By taking the partial derivatives of the likelihood function and setting them equal to zero, the likelihood equation for estimating the parameters is obtained. Thus we obtain

$$\frac{\partial LL}{\partial \beta_r} = \sum_{i=1}^{k} \left\{ (1 - d_i) I_{y_i>0} \left[\frac{g'_{\mu_i}}{g} - \frac{(1 + \alpha \mu_i)^{-\alpha^{-1}-1}}{1 - (1 + \alpha \mu_i)^{-\alpha^{-1}}} \right] \mu_i x_{ir} \right.$$
$$\left. + \frac{d_i}{\sum_{j=y_i}^{\infty} \Pr(Y_i = j)} \frac{\partial \sum_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \beta_r} \right\} = 0$$

$$\frac{\partial LL}{\partial \alpha} = \sum_{i=1}^{k} \left\{ (1 - d_i) I_{y_i>0} \left[\frac{g'_\alpha}{g} \right. \right.$$
$$\left. - \frac{\alpha^{-1} \mu_i / (1 + \alpha \mu_i) - \alpha^{-2} \log(1 + \alpha \mu_i)}{1 - (1 + \alpha \mu_i)^{-\alpha^{-1}}} (1 + \alpha \mu_i)^{-\alpha^{-1}} \right]$$
$$\left. + \frac{d_i}{\sum_{j=y_i}^{\infty} \Pr(Y_i = j)} \frac{\partial \sum_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \alpha} \right\} = 0$$

$$\frac{\partial LL}{\partial \delta_s} = \sum_{i=1}^{n} (1 - d_i) \left[I_{y_i=0} (1 - w_0) - I_{y_i>0} w_0 \right] z_{is} = 0$$

where

$$\frac{\partial \sum\limits_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \beta_r} = \sum_{j=y_i}^{\infty} (1 - w_0) \frac{g'_{\mu_i}(1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}) - (1 + \alpha\mu_i)^{-\alpha^{-1}-1}g}{(1 - (1 + \alpha\mu_i)^{-\alpha^{-1}})^2} \mu_i x_{ir}$$

$$\frac{\partial \sum\limits_{j=y_i}^{\infty} \Pr(Y_i = j)}{\partial \alpha} = \sum_{j=y_i}^{\infty} (1 - w_0) \left[\frac{g'_\alpha}{1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}} + \frac{\log(1 + \alpha\mu_i)\alpha^{-2} - \frac{\alpha^{-1}\mu_i}{1+\alpha\mu_i}}{\left(1 - (1 + \alpha\mu_i)^{-\alpha^{-1}}\right)^2} g(1 + \alpha\mu_i)^{-\alpha^{-1}} \right]$$

$$g'_\alpha = \left(\frac{\Gamma'(y_i + \alpha^{-1})}{\Gamma(y_i + \alpha^{-1})} - \frac{\Gamma'(\alpha^{-1})}{\Gamma(\alpha^{-1})} + \alpha^{-2}\log(1 + \alpha\mu_i) - \frac{\alpha^{-1} + y_i}{1 + \alpha\mu_i}\mu_i + \alpha^{-1}y_i \right) g$$

$$g'_\mu = \frac{y_i - \mu_i}{\mu_i(1 + \alpha\mu_i)} g$$

Since these partial derivative equations cannot be further simplified, we have applied the Newton-Raphson method with ridging as the optimization algorithm, using code in SAS given in the Appendix. Furthermore, the Convergence of the algorithm does not necessarily mean that a global maximum has been found, it just means that the convergence criteria have been achieved, and thus it can be a local maximum.

4. Goodness-of-fit statistics

For count regression models, a measure of goodness of fit may be based on the deviance statistic D defined as

$$D = -2 \left[\log L(\hat{\theta}_i; \hat{\mu}_i) - \log L(\hat{\theta}_i; y_i) \right] \tag{15}$$

where $\log L(\hat{\theta}_i; \hat{\mu}_i)$ and $\log L(\hat{\theta}_i; y_i)$ are the model's likelihood evaluated respectively under $\hat{\mu}_i$ and y_i. The log-likelihood functions are given in equation (7).

For an adequate model, the asymptotic distribution of the deviance statistic D is chi-square distribution with $n - k - 1$ degrees of freedom. Therefore, if the value for the deviance statistic D is close to the degrees of freedom, the model may be considered as adequate. When we have many regression models for a given data set, the regression model with the smallest value of the deviance statistic D is usually chosen as the best model for describing the given data.

In many data sets, the μ_i's may not be reasonably large and so the deviance statistic D may not be suitable. Thus, the log-likelihood statistic $\log(\hat{\theta}_i; y_i)$ can be used as an alternative statistic to compare the different models. Models with the largest log-likelihood value can be chosen as the best model for describing the data under consideration.

When there are several maximum likelihood models, one can compare the performance of alternative models based on several likelihood measures which have been proposed in the statistical literature. The AIC and BIC are two of the most regularly used measures. The AIC is defined as

$$AIC = -2l + 2p \qquad (16)$$

where l denotes the log likelihood evaluated for estimated parameters μ and p the number of parameters. For this measure, the smaller the AIC, the better the model is.

5. An application

The state wildlife biologists want to model how many fish[1] are being caught by fishermen at a state park. Visitors are asked how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there are no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish. We have data on 250 groups that went to a park. Each group was questioned about how many fish they caught (*count*), how many children were in the group (*child*), how many people were in the group (*persons*), and whether or not they brought a camper to the park (*camper*).

We will use the variables *child*, *persons* and *camper* in our model. Table 1 shows the descriptive statistics of using variables and also the *camper* variable has two values, zero and one as Table 2. In addition, Figure 1 shows the histogram of the *count* variable before censoring.

We have considered the model as follow

$$\log(\mu) = b_0 + b_1 * camper + b_2 * persons + b_3 * child,$$
$$\text{logit}(w_0) = a_0 + a_1 * child$$

Furthermore, we put two censoring points, $c_1 = 3, c_2 = 5$. Table 3 shows the estimation of the parameters according to different censoring constants. Also, the $-2\log L$ and AIC are presented as the goodness-of-fit measures.

Table 1: *Descriptive statistics for the fish data.*

Variable	Mean	Std Dev	Min	Max	Variance
count	3.296	11.635028	0	149	135.373880
child	0.684	0.850315	0	3	0.723036
persons	2.528	1.112730	1	4	1.238169

---

1. The fish dataset is available at the UCLA Academic Technology Services website, http://www.ats.ucla.edu.

Table 2: *camper variable description.*

camper	Frequency	Percent
0	103	41.2
1	147	58.8

Table 3: *Parameter estimation.*

Parameter	$c_1 = 3$	$c_2 = 5$
b_0	−1.0922	−0.9616
	(0.5998)	(0.4764)
b_1	0.7043	0.6079
	(0.3235)	(0.2702)
b_2	0.7397	0.7227
	(0.2086)	(0.1533)
b_3	−0.9130	−0.9266
	(0.3449)	(0.2807)
a_0	−0.3843	−0.3843
	(0.1703)	(0.1703)
a_1	1.1110	1.1110
	(0.2049)	(0.2049)
α	0.5673	0.6225
	(0.4388)	(0.3412)
$-2\log L$	540.9	618.1
AIC	554.9	632.1

According to the censoring points, there is 22.8% censored data when $c_1 = 3$. It means that 22.8% of the values of the response variable (*count*) are 0, 1, 2, 3 and the remaining 77.2% of values of the response variable are greater than 3, that is censored in the model. Also the percentage of the censoring for $c_2 = 5$ is 12%. For example, the 25th value of the response variable is $count_{25} = 30$, and the values of the independent variables are as follow

$$camper_{25} = 1, \ persons_{25} = 3, \ child_{25} = 0$$

So we want to censor only the value of the response variable ($count_{25} > censored\,point$).

The estimated parameter for camper variable of the model is a positive value for both censoring points, it means that while being a camper (*camper* = 1), the expected log(*count*) will be increased by 0.7043 and 0.6079, respectively when $c_1 = 3$ and $c_2 = 5$. Also, the effect of *persons* is positively associated with the number of fish caught for both censoring points, and the expected log(*count*) will be increased for a unit increase in *persons* for the first and the second censoring point, respectively, by 0.7397 and 0.7227. But, the expected log(*count*) will be decreased for a unit increase in *child* for by 0.9130 and 0.9266, respectively for the first and the second censoring point. Further-
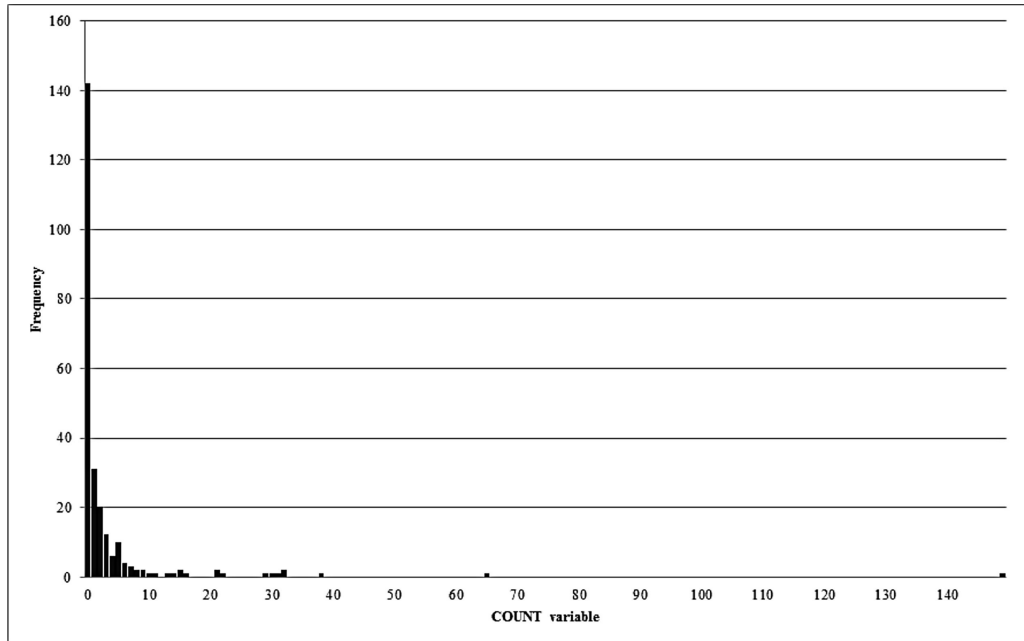
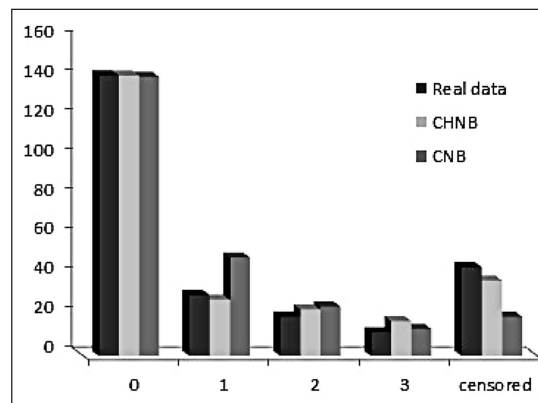***Figure 1:*** *Histogram of the response variable.*

Figure 2: *Estimate vs Real when the censoring point is 3.*

more, the estimated parameter for *child* variable of the logit part of the model is positive for both censoring points, i.e, the group with more children was less successful in fishing. The estimated dispersion parameter suggested overdispersion in the model for both censoring points.

We have compared the censored hurdle negative binomial (CHNB) regression model with the censored negative binomial (CNB) regression model and the results are shown in Figure 2 and Figure 3 when the censoring points are $c_1 = 3$, $c_2 = 5$, respectively. The CHNB regression model shows a better estimation than the CNB regression model except when the value of the count variable is 3 in Figure 2 and 4 in Figure 3. The CHNB
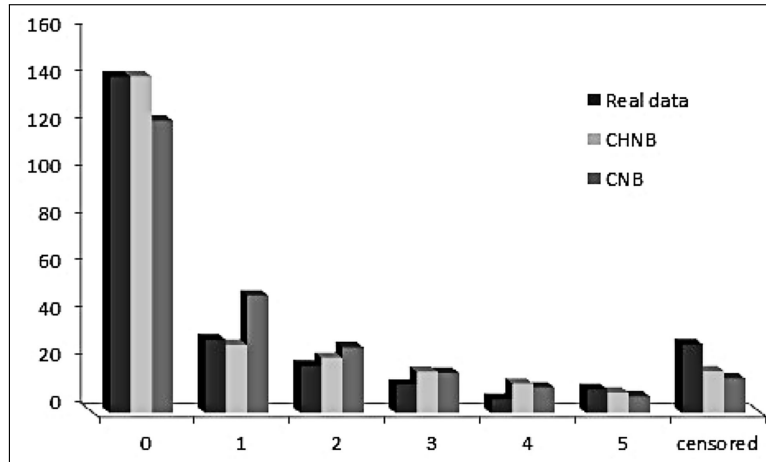
Figure 3: *Estimate vs Real when the censoring point is 5.*

regression model estimated the number of zeros as 142 (the same as number of zeros in fish data) in both cases ($c_1 = 3, c_2 = 5$), but the CNB regression model estimated 141 and 123, respectively, when the censoring points are $c_1 = 3, c_2 = 5$. Also, the CHNB regression model shows a closer estimate for the censored values than the CNB regression model. For example, when there are 45 censored values in the fish data, the CHNB regression model presented 38 censored values and the CNB regression model presented 20 censored values.

6. Conclusion

In this article we want to show that the hurdle negative binomial regression model can be used to fit right censored data. In fact, the proposed model is suitable to solve the excess zeros problem in the response variable when the data are censored from the right side. The results from the fish data are summarized in Table 1-3. The goodness-of-fit measures are presented in the Table 3 according to different censoring points and it is obvious that we have a smaller value for $-2\log L$ or AIC when the percentage of censoring increase and that is because of the number of the data which are used in the model. Also, the censored hurdle negative binomial model shows a better fit with respect to the censored negative binomial model for different censoring points as shown in Figure 2 and Figure 3.

Acknowledgment

References

Arulampalam, W. and Booth, A. (1997). Who gets over the training hurdle? a study of the training experiences of young men and women in britain. *Journal of Population Econometrics*, 10, 197–217.

Boucher, J. P. and Guillen, M. (2009). A survey on models for panel count data with applications to insurance. *RACSAM, Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas*, 103(2), 277–294.

Chou, N. T. and Steenhard, D. (2009). A Flexible Count Data Regression Model Using SAS. *PROC NLMIXED. Proceedings SAS Global Forum*, paper 250-2009.

Dalrymple, M., Hudson, I. and Barnett, A. (2003). Finite mixture, zero-inflated poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, 41, 491–504.

Dionne, G. and Vanasse, C. (1989). A generalization of automobile insurance rating models: the negative binomial distribution with regression component. *Astin Bulletin*, 19, 199–212.

Greene, W. (2005). Functional form and heterogeneity in models for count data. *Foundations and Trends in Econometrics*, 1(2), 113–218.

Grootendorst, P. V. (1995). A comparison of alternative models of prescription drug utilization. *Health Economics*, 4, 183–198.

Gurmu, S. and Trivedi, P. K. (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics*, 14, 469–477.

Gurmu, S. (1998). Generalized hurdle count data regression models. *Economics Letters*, 58, 263–268.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.

Pohlmeier, W. and Ulrich, V. (1995). An econometric model of the two-part decision-making process in the demand for health care. *The Journal of Human Resources*, 30, 339–361.

Saffari, S. E. and Robiah, A. (2011a). Zero-inflated poisson regression models with right censored count data. *Matematika*, 27(1), 21–29.

Saffari, S. E. and Robiah, A. (2011b). Zero-inflated negative binomial regression model with right censoring count data. *Journal of Materials Science and Engineering*, B 1, 551–554.

Saffari, S. E., Robiah, A. and Greene, W. (2011). Handling of over-dispersion of count data via truncation using poisson regression model. *Journal of Computer Science and Computational Mathematics*, 1(1), 1–4.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307–333.

Winkelmann, R. (2003). Health care reform and the number of doctor visits – an econometric analysis. *Journal of Applied Econometrics*, 19, 455–472.

Appendix:
SAS code to estimate the parameters for Fish data

```
data fish;
set fish;
bound=3;
if count > bound then count=bound+1; * This is probably
 how you would see the data if it was actually censored;

proc nlmixed TECH=NRRIDG;
parms a_0=-0.4 a_1=1 b_0=-1 b_1=1 b_2=1 b_3=-1 alpha=0.5;
bounds alpha>0;
lin = a_0 + a_1* child;
w = exp(lin)/(1+exp(lin));
eta = b_0 + b_1 *camper + b_2* persons + b_3* child;
   mu = exp(eta);
phi=1/alpha;
pdf=(gamma(count+phi)/(gamma(count+1)*gamma(phi)))
*((1/(1+alpha*mu))**phi*(alpha*mu/(1+alpha*mu))** count);
l_1 =  w;
l_2 = (1-w) * pdf / (1-(1+alpha*mu)**(-phi));
cdf=0;
do t=1 to bound;
cdf=cdf+(gamma(t+phi)/(gamma(t+1)*gamma(phi)))
*((1/(1+alpha*mu))**phi*(alpha*mu/(1+alpha*mu))**t);
end;
l_3= (1-w)*(1-cdf/(1-(1+alpha*mu)**(-phi)));
if count = 0 then ll = log(l_1);
if  0 < count <= bound then ll = log(l_2);
if count <= bound then d=0; else d=1;
      ll=(1-d)*ll+d*log(l_3);
model count ~general(ll);
predict mu out=hnbmu;
predict w out=hnbw;
run;
```