Journal of Economics and Behavioral Studies (ISSN: 2220-6140)
Vol. 9, No. 3, pp. 6-18, June 2017

The Effect of Sample Size on the Efficiency of Count Data Models: Application to Marriage Data

Volition Tlhalitshi Montshiwa, Ntebogang Dinah Moroke
North West University, South Africa
volition.montshiwa@nwu.ac.za, ntebo.moroke@nwu.ac.za

Abstract: Sample size requirements are common in many multivariate analysis techniques as one of the measures taken to ensure the robustness of such techniques, such requirements have not been of interest in the area of count data models. As such, this study investigated the effect of sample size on the efficiency of six commonly used count data models namely: Poisson regression model (PRM), Negative binomial regression model (NBRM), Zero-inflated Poisson (ZIP), Zero-inflated negative binomial (ZINB), Poisson Hurdle model (PHM) and Negative binomial hurdle model (NBHM). The data used in this study were sourced from Data First and were collected by Statistics South Africa through the Marriage and Divorce database. PRM, NBRM, ZIP, ZINB, PHM and NBHM were applied to ten randomly selected samples ranging from 4392 to 43916 and differing by 10% in size. The six models were compared using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Vuong's test for over-dispersion, McFadden RSQ, Mean Square Error (MSE) and Mean Absolute Deviation (MAD).The results revealed that generally, the Negative Binomial-based models outperformed Poisson-based models. However, the results did not reveal the effect of sample size variations on the efficiency of the models since there was no consistency in the change in AIC, BIC, Vuong's test for over-dispersion, McFadden RSQ, MSE and MAD as the sample size increased.

Keywords: *Poisson regression, Negative binomial regression, Zero-inflated Poisson, Zero-inflated negative binomial, Poisson Hurdle and Negative binomial hurdle*

1. Introduction

Count data is defined by Hilbe (2014)as observations that only take non-negative integers theoretically ranging from zero to the maximum value of the variable being modelled. Poisson regression model (PRM) is used as the basis for modelling count responses under the assumption that the conditional mean of the outcome variable is equal to the conditional variance (equi-dispersion) (Vach, 2012). However, as much as this is a naturally occurring basic property of the Poisson distribution, it is not always true in real life datasets and count response data may exhibit under-/over-dispersion (SAS-Institute, 2012; Tang et al., 2012; Vach, 2012).These authors cautioned that violation of the equi-dispersion assumption results in inefficient, potentially biased parameter estimates and small standard errors of the PRM. As such, SAS-Institute (2012) recommends the negative binomial regression model (NBRM) as an extension of PRM in situations where the variance is significantly bigger than the conditional mean (over-dispersion).

A limitation of both the PRM and NBRM occurs when there are too many zeroes (excess-/ extra-zeroes) in the count outcome variable. This may be due to either non-response (structural or unobserved zeros) or many respondents having a count of zero for the outcome variable being measured (observed zeros) (Little, 2013; Wang, Xie, Fisher & Press, 2011). Excess-zeroes in the count outcome variable may distort the expectation and variance values of some covariates when PRM and/or NBRM are used for modelling such count data (Little, 2013). As such, the zero-inflated Poisson (ZIP) model was designed to model count response data with excess zeros when the assumption of equi-dispersion holds (SAS-Institute, 2012). On other hand, zero-inflated negative binomial (ZINB) was formed to model over-dispersed count response data with excess zeros (SAS-Institute, 2012). Other challenges that may arise in count response modelling are under-dispersion and zero-deflation but they seldom occur in practice (Morel & Neerchal, 2012; Ozmen & Famoye, 2007). Despite their seldom occurrence in practice, under-dispersion and zero-deflation have led to the birth of hurdle models namely: the Poisson Hurdle model (PHM) and Negative Binomial Hurdle model (NBHM) which are described in detail by Rose, Martin, Wannemuehler & Plikaytis (2006).The models considered in this study based on their popularity are: PRM, NBRM, ZIP, ZINB, PHM and NBHM.

2. Literature Review

Literature shows that several studies compared numerous count data models using different datasets. There is evidence that count data models are evolutional in that previous research worked towards developing models that can remedy the shortcomings of the existing ones. However, there is no count data model that has been found to be generally ideal. Several authors including Ver Hoef and Boveng (2007) and Rose et al. (2006) caution that the choice of the model is dependent on the theoretical and/or scientific knowledge of the data being modelled. Most authors such as Burger, Van Oort and Linders (2009), Famoye and Singh (2006), Mei-Chen, Pavlicova and Nunes (2011), Rose et al. (2006) and, Yip and Yau (2005) compared PRM, NBRM, ZIP, ZINB, PHM and NBHM to other count data models including the quasi-Poisson regression model (QPRM), zero-inflated generalized Poisson (ZIGP) and zero inflated double- Poisson (ZIDP). It is therefore evident that PRM, NBRM, ZIP, ZINB, PHM and NBRM are the most commonly used count data models in literature hence the scope of this study is limited to these six models. The most common criteria for comparing count data models in literature are AIC, Vuong test, goodness of fit tests and generalised Pearson's Chi-square (Burger et al., 2009; Ver Hoef & Boveng 2007; Mei-Chen et al., 2011; Rose et al., 2006; Yip & Yau, 2005). These criteria are adopted in the current study and are discussed in detail in the methodology section. The current study acknowledges that the efficiency of count data models is mainly affected by poor data quality (excess zeroes) and violations of distributional assumptions (under-/ over-dispersion for instance) hence effort should be made to improve on data quality rather than just re-parameterisation of count data models.

Despite the common practice of sample size considerations in multivariate analysis, many previous studies around the application of count data models have not focused on sample size considerations. More specifically, literature shows that most studies have only compared various count data models under one sample size (Fuzi, Jemain & Ismail, 2016; Ver Hoef & Boveng, 2007; Mei-Chen et al., 2011; Park, Lord & Hart, 2010; Rose et al., 2006). As such, this study largely seeks to understand whether or not sample size variations can improve the efficiency of count data models relative to under/ over-dispersion and excess zeroes without further iterative re-parameterisation of the known models with the intent to bridge a gap in literature around the application of count data models. Another motivation for conducting this study is that marriage and divorces datasets usually have both metric and categorical variables but previous studies did not apply count data models to such data despite their ability to model both metric and categorical variables. As such, this study contributes a new idea to the research around marriage and divorces by applying count data models to such data. In essence, the study compares the efficiency of the most commonly used count data models under different sample sizes.

3. Methodology

Description of data: The data used in this study were sourced from Data First, available at https://www.datafirst.uct.ac.za. The proposed datasets are for the periods of 2010 (N=22936) and 2011 (N=20980). These data sets are used because they are readily available and also share the methodology of collection as opposed to other datasets collected prior to 2010. The data were collected by Statistics South Africa (StasSA) using a standard structured form (Divorce Form 07-04) prepared by StatsSA in collaboration with the Department of Justice. The categorical variables used in this study are: Male Race, Female Race, Male Occupation, Female Occupation, Male Status (Marital status of husband), Female Status (Marital status of wife), Male No Times Married, Female No Times Married, Solemnisation, and Marriage Type. The continuous variables are Male Age, Female Age, No of Children and Duration of Marriage (dependent variable). The choice of these variables is embedded on literature which has identified these selected socio-demographic variables of the couples as significant predictors of marriage life (Cox & Demmitt, 2013; Holman, 2006; Reis & Sprecher, 2009).

It is worth noting that the main focus of this paper is on the application of models and not on the prediction of the Duration of Marriage. Data are analysed using the Statistical Analysis Software (SAS®) version 9.3, registered to the SAS Institute Inc. Cary, NC, USA. The datasets for 2010 and 2011 were merged using the MERGE statement of SAS® as recommended by Tilanus (2008) to form a total of N=43916. Since the interest of the current study is to explore the performance of count data models under different sample sizes, the merged dataset for 2010 and 2011 (N=43916) were further divided into ten random samples in multiples of

10% until 100% (N). The random sampling was performed using the SQL procedure of SAS® as recommended by Matignon (2007). There is no specific theoretical reason for choosing these sample sizes (at 10% increments) but the general intention is to simulate scenarios in which count data models are applied to different sample sizes. The samples are randomly selected from the merged dataset with the intention of preserving the distributional characteristics of the merged data (mean and variance, See Table 1). By keeping the means and variances similar across the samples, one may ensure that the severity of under-/ over-dispersion does not differ much across the samples hence the effect of sample size on the proposed models can be assessed more accurately. The general intent here is to minimise the hallo effects and ensure that the aspect that significantly differentiate between the samples is sample size, which is the main interest of this study. An advantage for drawing samples from a real life dataset is that the samples depict the real life scenarios better than the completely simulated datasets.

Generalised Linear Models (GLMs): Zeileis, Kleiber and Jackman (2008) emphasise that all count data models belong to the Generalised Linear Models (GLM) family. The authors also explained that all GLMs use the same log-linear mean function which is defined in (1) but make different assumptions about the remaining likelihood. The general GLM is defined by:

$$\log \mu = x^T \beta, \qquad (1)$$

where μ, x^T and β denote the mean, the transpose of a vector of regressors and the parameter vector respectively. For all count data models under study, the ML algorithm is used in maximising the $\mathcal{L}$ function to enable the estimation of parameter estimates. This ML algorithmic known as the Newton Raphson (SAS Institute, 2010). The algorithm updates the parameter vector $\boldsymbol{\beta_r}$at each iteration with (2):

$$\boldsymbol{\beta_{r+1}} = \boldsymbol{\beta_r} - \boldsymbol{H^{-1}s}, \qquad (2)$$

where $\boldsymbol{s}$ is the gradient or score matrix generated from the first derivative of $\mathcal{L}$ and $\boldsymbol{H}$is the Hessian Matrix generated from the second derivative of $\mathcal{L}$ at the current value of the parameter. More specifically, $\boldsymbol{s}$and $\boldsymbol{H}$are computed using (3) and (4) respectively:

$$\boldsymbol{s} = [s_j] = \left[\frac{\partial \mathcal{L}}{\partial \beta_j}\right] = 0 \qquad (3)$$

$$\boldsymbol{H} = [h_{ij}] = \left[\frac{\partial^2 \mathcal{L}}{\partial \beta_i \partial \beta_j}\right] = 0 \qquad (4)$$

Poison and negative binomial models: For PRM, the current study adopts the methods explained by Karlaftis, Mannering and Washington (2010) unless otherwise specified. The probability of the count outcome variable$\boldsymbol{y_i}$ is given by (5):

$$P(y_i|x_i) = \frac{e^{(-\lambda_i)\lambda_i^{y\,i}}}{y_i!}, \qquad (5)$$

where x_i denote the ith set of predictors per couple. The parameter λ_i is defined by the expected value of y_i.The parameters for PRM are estimated by maximising the Poisson log- likelihood ($\mathcal{L}$) function expressed in (6) (Hilbe, 2014):

$$\mathcal{L}(\beta : y_i) = \sum_{i=1}^{n} \left\{ y_i \ln\left(x_i' \beta\right) - e^{\left(x_i' \beta\right)} - \ln y_i \right\} \qquad (6)$$

NBRM is designed to be used when there is significant over-dispersion in the data and accounts for over-dispersion by including an extra parameter in the PRM model (the dispersion parameter). The general probability function for NBRM is defined by (7) and is adopted from Whitehead, Haab and Huang (2012) as:

$$P(y_i|x_i) = \frac{\Gamma\left(y_i+\alpha^{-1}\mu_i^{2-p_i}\right)\alpha^{y_i}\mu_i^{(p_iy_i-2y_i)}\left(1+\alpha\mu_i^{p_i-1}\right)^{-\left(y_i+\alpha^{-1}\mu_i^{2-p_i}\right)}}{\Gamma(y_i+1)\Gamma\left(\alpha^{-1}\mu_i^{2-p_i}\right)}, \tag{7}$$

where Γ is the Gamma function,

$$\mu_i = E(y_i|\boldsymbol{\beta}), \tag{8}$$

which follows a binomial probability distribution, p_i and α are additional parameters that allow for flexibility in over-dispersion. The parameter estimates for NBRM are obtained by maximising the $\mathcal{L}$ function in (9) (adopted from (Hilbe, 2014) :

$$\mathcal{L}(\mu; y, \alpha) = \sum_{i=1}^{1} y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i+\frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right). \tag{9}$$

Zero inflated models: ZIP is a special form of PRM which is used when the equi-dispersion assumption holds but the count outcome variable exhibits zero-inflation. The equations discussed in this section are adopted from the SAS Institute (2010) unless otherwise specified. The probability density function of $y$for ZIP is given by (10):

$$f(y) = \begin{cases} \omega + (1-\omega)e^{-\lambda}, for\ y = 0 \\ (1-\omega)\frac{\lambda^y e^{-\lambda}}{y!}, for\ y = 1,2,3\ldots \end{cases}, \tag{10}$$

where ω denotes the zero-inflation probability and λ is the Poisson mean parameter. The parameter estimates for ZIP are obtained by maximising the $\mathcal{L}$ function in (11) using the ML algorithm.

$$\mathcal{L} = \begin{cases} w_i \log\left[\omega_i + (1-\omega_i)e^{-\lambda}\right], for\ y_i = 0 \\ w_i \log[(1-\omega_i) + y_i \log(\lambda_i) - \lambda_i - \log(y_i!)], for\ y_i > 0 \end{cases}, \tag{11}$$

where w_i denotes the weight of the observation.

The probability density functions of $y$underZINB are given by (12):

$$f(y) = \begin{cases} \omega + (1-\omega)(1+k\lambda), for\ y = 0 \\ (1-\omega)\frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)}\frac{(k\mu)^y}{(1+k\lambda)^{y+1/k}},\ for\ y = 1,2,3\ldots \end{cases}, \tag{12}$$

where k is the negative binomial dispersion parameter and ω denotes the zero-inflation probability. Parameter estimates for ZINB are obtained by maximising the $\mathcal{L}$ function in (13) which is defined by:

$$\mathcal{L} = \begin{cases} \log\left[\omega_i + (1-\omega_i)\left(1+\lambda\frac{k}{\omega_i}\right)\right], for y_i = 0 \\ \log(1-\omega_i) + y_i \log\left(\frac{k\lambda}{\omega_i}\right) \\ -\left(y_i + \frac{\omega_i}{k}\right)\log\left(1+\frac{k\lambda}{\omega_i}\right) \\ + \log\left(\frac{\Gamma\left(y_i+\frac{\omega_i}{k}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{\omega_i}{k}\right)}\right), for y_i > 0 \end{cases}. \tag{13}$$

SAS Institute (2010) elaborate that there are two link functions and linear predictors associated with zero-inflated distributions of which one is for the zero inflation probability (ω) and another is for the mean parameter (λ).

Hurdle models: Hurdle models are designed to address both under/ over-dispersion. In order to obtain the zero-truncated forms of the probability density functions (PDF's), this study adopts the methods explained by Stroup (2012). The general form of the zero-truncated Poisson PDF corresponding to PHM is given by (14):

$$P(Y) = \frac{e^{(-\lambda_i)}\lambda_i^y}{y_i!(1-e^{-\lambda_i})}. \tag{14}$$

The parameter estimates for PHM are obtained by maximising the $\mathcal{L}$ function in (15) using the ML algorithm.

$$\mathcal{L} = \begin{cases} \log p_i \quad for\ y=0 \\ \log(1-p_i)y\log\lambda_i - \lambda_i - \log(y!) - \log\left[1-e^{-\lambda_i}\right] \quad for\ y = 1,2,\ldots \end{cases} \tag{15}$$

The PDF for the zero-truncated negative binomial model corresponding to NBHM is defined by (16):

$$P(Y) = \frac{\binom{y+\left(\frac{1}{\alpha}\right)-1}{y}\left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^y\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha}}{1-\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha}}. \tag{16}$$

In order to obtain the parameter estimates, the $\mathcal{L}$ function of NBHM in (17) is maximised using the ML algorithm.

$$\mathcal{L} = \begin{cases} \log \mathrm{p_i} \quad \text{for y=0} \\ \log(1-p_i) + y\log\frac{\alpha\lambda_i}{1+\alpha\lambda_i} - \frac{1}{\alpha}\log(1+\alpha\lambda_i) + \log\left\{\binom{y+\left(\frac{1}{\alpha}\right)-1}{y}\right\} \quad for\ y = 1,2,\ldots \end{cases} \tag{17}$$

Model Comparison Criteria: The comparison of the six proposed count data models considered in this study was done at two phases namely: the within-sample comparison and the between-sample comparison stage. The within-sample comparison phase entails comparing the proposed models within a specific sample size and selecting the most efficient model based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), both recommended by Hilbe (2014). The Vuong's test for over-dispersion (Rose et al., 2006; Little, 2013) and the McFadden's RSQ (Karlaftis et al., 2010) are also used as other criteria to avoid biasness. The model that minimises the values of AIC and BIC but maximises the McFadden's RSQ is preferred (Hilbe, 2014; Karlaftis et al., 2010). The Vuong's test and LRT generally test the hypothesis that the dispersion parameter is zero. The values of the Vuong statistic less than -1.96 favours the null model while values greater than 1.96 favours the proposed alternative (Alt) model whereas the $|V| < 1.96$ yields inconclusive results (Peng, Shi, Nagaraja & Xiang, 2014).

The six models that were selected from each sample were then compared to each other in the between-sample comparison phase with the aim of selecting the most effective model from the proposed six. The McFadden'sRSQ (Karlaftis et al., 2010), the mean absolute deviation (MAD) suggested by Bajpai (2009) and the mean squared error (MSE) adopted from Park et al. (2010). The model which maximises the McFadden's RSQ is preferred (Karlaftis et al., 2010) whereas the model which minimises the MAD and MSE is preferred (Bajpai, 2009 & Park et al., 2010). The equations of model comparison criteria used in both the within- and between-sample comparisons are given in Appendix 1. Following the selection of the most effective model is the test for overall significance of the model using the likelihood ratio chi-square test (Berk & Carey, 2009) and the significance of each parameter estimate in the selected model using the Wald's test (Cameron & Trivedi, 2013).

4. Data Analysis and Results

Dispersion and zero-inflation: Since the under-/ over-dispersion and excess zeros are of interest when using count data models, this study analysed the percentage of zeros and the measures of dispersion (mean and variance) as presented in Table 1.

Table 1: Sample sizes, percentage of zeros and dispersion of Duration of Marriage

Sample	N	Mean	Variance	% 0's
10%	4392	11.09	74.523	4.94
20%	8783	11.164	74.89	4.94
30%	13175	11.139	74.541	4.93
40%	17566	11.135	74.781	5.04
50%	21958	11.108	74.422	5.16
60%	25107	11.612	71.421	5.17
70%	29311	11.604	71.367	5.09
80%	33478	11.592	71.339	5.09
90%	37667	11.591	71.228	5.14
100%	41881	11.589	71.153	5.08

Table 1 shows that the mean and variance of Duration of Marriage are approximately equal across all samples. The variance is about seven times more than the mean implying that the dependent variable (Duration of Marriage) may be over-dispersed. The study therefore implemented Vuong's test for over-dispersion in order to confirm whether significant over-dispersion exists in the data or not (See Table 2). The percentage of zeroes in each sample is approximately five implying that the Duration of Marriage is not zero-inflated for all sample sizes. The results in Table 1 suggest that NBRM which is theoretically designed to analyse over-dispersed data that are not zero-inflated (SAS-Institute, 2012) is appropriate for the data sets under study. However, for experimental purposes, this study fitted all the six proposed models, compared them and selects the most effective model under different sample sizes.

Within-sample comparison phase: Table 2 shows that at Stage 1, PRM is compared to ZIP and it is evident that all criteria are in favour of ZIP except for the 50% sample size where McFadden's RSQ favoured PRM. Generally, Stage 1 favoured ZIP over PRM. The Vuong's test favours ZIP over PRM indicating the need to model zero-inflation in the data under the assumption of equi-dispersion. Based on the collective results in Stage 1, PRM is eliminated from the comparison and ZIP is compared to NBRM at Stage 2. Highlighted in Stage 2 is that ZIP has a bigger McFadden's RSQ but the majority of criteria are in favour of NBRM. The dispersion parameter (alpha) was significantly greater than zero confirming that the data are over-dispersed (Liu and Cela, 2008). Since NBRM was found to be more appropriate for modelling the Duration of Marriage at Stage 2, this model was compared to ZINB in Stage 3 which is proved to be by all selection criteria. As such, ZINB is compared to PHM in Stage 4 where all but one comparison statistics is favouring ZINB. Stage 5 compares ZINB to NBHM and the AIC and BIC are in favour of NBHM but ZINB slightly outperforms NBHM in terms of the variation explained in the Duration of Marriage by the predictor (Mc Fadden's RSQ). Due to its inconclusive results ($|V| < 1.96$) in comparing NBHM and other models using the Vuong's test, the results of this test were not stated in Table 2 and Table 3 but AIC, BIC and McFadden's RSQ are reported. NBHM is chosen as the best model for fitting the Duration of Marriage for the 10%, 20%, 30%, 40% and 50% sample sizes and is compared to other models in the within-sample comparison phase. The results for the 10% to the 50% sample sizes are collated into Table 1 because they lead to the same conclusion about the preferred model. Similarly, the results for sample sizes of 60% to 100% are collated into Table 2 for the same reasons.

Table 2: Within-sample comparison and selection (10% to 50% sample sizes)

				10%	**20%**	**30%**
STAGE	**NULL**	**ALT**	**CRITERION**	**COMPARISON STATISTICS**	**COMPARISON STATISTICS**	**COMPARISON STATISTICS**
1	PRM	ZIP	VUONG	V(4.766752)>1.96	V(7.408155)>1.96	V(9.3474)>1.96
			AIC	PRM(13444)>ZIP(13178)	PRM(27508)>ZIP(26825)	PRM(41193)>ZIP(39956)
			BIC	PRM(13518)>ZIP(13325)	PRM(27590)>ZIP(26990)	PRM(41280)>ZIP(40131)
			McFadden RSQ	PRM(0.70869)<ZIP(0.932)	PRM(0.70200)>ZIP(0.688)	PRM(0.701474)<ZIP(0.711)
2	ZIP	NBRM	VUONG	V(3.54722)>1.96	V(6.34488)>1.96	V(7.61848)>1.96
			AIC	ZIP(13178)>NBRM(12349)	ZIP(26825)>NBRM(25057)	ZIP(39956)>NBRM(3734)
			BIC	ZIP(13325)>NBRM (12429)	ZIP(26990)>NBRM (25146)	ZIP(40131)>NBRM (37434)

			McFadden RSQ	ZIP(0.6937)>NBRM(0.587)	ZIP(0.702)>NBRM(0.582)	ZIP((0.711)>NBRM(0.584)
3	NBRM	ZINB	VUONG	V (114.775)>1.96	V (83.64964)>1.96	V (100.218)>1.96
			AIC	NBRM(12349)>ZINB(12290)	NBRM(25057)>ZINB(24872)	NBRM(373397)>ZINB(36994)
			BIC	NBRM(12429)>ZINB(12442)	NBRM(25146)>ZINB(25044)	NBRM(25146)>ZINB(37176)
			McFadden RSQ	NBRM(0.58744) <ZINB(0.59029)	NBRM(0.58162) <ZINB(0.58514)	NBRM(0.58382) <ZINB(0.58796)
4	ZINB	PHM	VUONG	V(-4.39533)<-1.96	V(-7.54295)<-1.96	V(-9.194521)<-1.96
			AIC	ZINB(12290)<PHM(13179)	ZINB(24872)<PHM(26826)	ZINB(36994)<PHM(39954)
			BIC	ZINB(12442)<PHM(13326)	ZINB(25044)<PHM(26991)	ZINB(37176)<PHM(40129)
			McFadden RSQ	ZINB(0.59029) <PHM(0.69367)	ZINB(0.58514) <PHM(0.68773)	ZINB(0.58514) <PHM(0.688825)
5	ZINB	NBHM	AIC	ZINB(12290)>NBHM(12288)	ZINB(24872)>NBHM(24868)	ZINB(36994)>NBHM(36990)
			BIC	ZINB(12442)>NBHM(12441)	ZINB(25044)>NBHM(25039)	ZINB(37176)>NBHM(37172)
			McFadden RSQ	ZINB(0.59029)>NBHM(0.5895)	ZINB(0.58514)<NBHM(0.58426)	ZINB(0.58796)>NBHM(0.58704)

Table 2: Within-sample comparison and selection (10% to 50% sample sizes) continued

				40%	50%	
STAGE	**NULL**	**ALT**	**CRITERION**	**COMPARISON STATISTICS**	**COMPARISON STATISTICS**	**PREFERRED MODEL**
1	PRM	ZIP	VUONG	V(11.274)>1.96	V(13.504)>1.96	ZIP
			AIC	PRM(59710)>ZIP(57632)	PRM(83309)>ZIP(80785)	ZIP
			BIC	PRM(59803)>ZIP(57817)	PRM(83403)>ZIP(80974)	ZIP
			McFadden RSQ	PRM(0.67619)>ZIP(0.663)	PRM(0.639)>ZIP(0.622)	ZIP
2	ZIP	NBRM	VUONG	V(9.374)>1.96	V(39.780)>1.96	NBRM
			AIC	ZIP(57632))>NBRM(53831)	ZIP(80785)>NBRM(67384)	NBRM
			BIC	ZIP(57817)>NBRM (53931)	ZIP(80974)>NBRM (67486)	NBRM
			McFadden RSQ	ZIP(0.663)>NBRM(0.550)	ZIP(0.6219)>NBRM(0.549)	ZIP
3	NBRM	ZINB	VUONG	V (113.220)>1.96	V (114.775)>1.96	ZINB
			AIC	NBRM(53831)>ZINB(53241)	NBRM(67384)>ZINB(67240)	ZINB
			BIC	NBRM(53931)>ZINB(53433)	NBRM(67486)>ZINB(67436)	ZINB
			McFadden RSQ	NBRM(0.550) <ZINB(0.555)	NBRM(0.549) <ZINB(0.550)	ZINB
4	ZINB	PHM	VUONG	V(-11.210)<-1.96	V(-42.294)<-1.96	ZINB
			AIC	ZINB(53241)<PHM(57630)	ZINB(67240)<PHM(80768)	ZINB
			BIC	ZINB(53433)<PHM(57814)	ZINB(67436)<PHM(80956)	ZINB
			McFadden RSQ	ZINB(0.555) <PHM(0.664)	ZINB(0.550) <PHM(0.622)	PHM
5	ZINB	NBHM	AIC	ZINB(53241)>NBHM(51492)	ZINB(67240)>NBHM(67207)	NBHM
			BIC	ZINB(53433)>NBHM(51591)	ZINB(67436)>NBHM(67402)	NBHM
			McFadden RSQ	ZINB(0.555)>NBHM(0.554)	ZINB(0.550) >NBHM(0.549)	ZINB

Table 3: Within-sample comparison and selection (60% to 100% sample sizes)

				60%	70%	80%
STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	COMPARE STATISTICS	COMPARE STATISTICS
1	PRM	ZIP	VUONG	V(2.924)>1.96	V(2.958)>1.96	V(2.942)>1.96
			AIC	PRM(77709)>ZIP(77500)	PRM(91181)>ZIP(90958)	PRM(104421)>ZIP(104181)
			BIC	PRM(77805)>ZIP(77693)	PRM(91280)>ZIP(91155)	PRM(104521)>ZIP(104382)
			McFadden RSQ	PRM(0.687)>ZIP(0.686)	PRM(0.685)>ZIP(0.684)	PRM(0.684)>ZIP(0.684)
2	ZIP	NBRM	VUONG	V(4.92629)>1.96	V(4.92629)>1.96	V(6.36931)>1.96
			AIC	ZIP(77500)>NBRM(71850)	ZIP(90958)>NBRM(84248)	ZIP((104181)>NBRM(96325)
			BIC	ZIP(77693)>NBRM (71953)	ZIP(91155)>NBRM(84355)	ZIP(104382)>NBRM(96433)
			McFadden RSQ	ZIP(0.686)>NBRM(0.577)	ZIP(0.684)>NBRM(0.575)	ZIP(.68356)<NBRM(0.575)
3	NBRM	ZINB	AIC	NBRM(71850)	NBRM(84248)	NBRM(96325)
			BIC	NBRM(719536)	NBRM(84355)	NBRM(96433)
			McFadden RSQ	NBRM(0.577)	NBRM(0.575)	NBRM(0.575)
4	NBRM	PHM	VUONG	V(-4.783)<-1.96	V(-5.093)<-1.96	V(-6.201)<-1.96
			AIC	NBRM(71850)<PHM(77481)	NBRM(84248)<PHM(90938)	NBRM(96325)<PHM(104154)
			BIC	NBRM(719536)<PHM(71233)	NBRM(84355)<PHM(91135)	NBRM(96433)<PHM(104354)
			McFadden RSQ	NBRM(0.577) <PHM(0.686)	NBRM(0.575) <PHM(0.684)	NBRM(0.575) <PHM(0.684)
5	NBRM	NBHM	AIC	NBRM(71850)<NBHM(73539)	NBRM(84248)<NBHM(86277)	NBRM(96325)<NBHM(98670)
			BIC	NBRM(719536)>NBHM(73740)	NBRM(84355)<NBHM(86482)	NBRM(96433)<NBHM(98878)
			McFadden RSQ	NBRM(0.57710)<NBHM(0.58866)	NBRM(0.57518)<NBHM(0.586296)	NBRM(0.57467) <NBHM(0.586)

Table 3: Within-sample comparison and selection (60% to 100% sample sizes) Continued

				90%	100%	
STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	COMPARE STATISTICS	PREFERRED MODEL
1	PRM	ZIP	VUONG	V(3.202)>1.96	V(3.729)>1.96	ZIP
			AIC	PRM(117481)>ZIP(117200)	PRM(293962)>ZIP(292859)	ZIP
			BIC	PRM(117583)>ZIP(117403)	PRM(294072)>ZIP(293080)	ZIP
			McFadden RSQ	PRM(0.684)>ZIP(0.683)	PRM(0.289)>ZIP(0.292)	PRM
2	ZIP	NBRM	VUONG	V(6.420)>1.96	V(38.702)>1.96	NBRM
			AIC	ZIP(117403)>NBRM(108393)	ZIP(293962)>NBRM(240502)	NBRM
			BIC	ZIP(104382)>NBRM(108503)	ZIP(294072)>NBRM(240622)	NBRM
			McFadden RSQ	ZIP(0.683)>NBRM(0.575)	ZIP(0.289)>NBRM(0.151)	ZIP
3	NBRM	ZINB	AIC	NBRM(108393)	NBRM(240502)	Second-order optimality condition violated.
			BIC	NBRM(108503)	NBRM(240622)	
			McFadden RSQ	NBRM(0.575)	NBRM(0.151)	
4	NBRM	PHM	VUONG	V(-6.292)<-1.96	V(-38.393)<-1.96	NBRM
			AIC	NBRM(108393)<PHM(117180)	NBRM(240502)<PHM(292778)	NBRM
			BIC	NBRM(108503)<PHM(117383)	NBRM(240622)<PHM(292999)	NBRM
			McFadden RSQ	NBRM(0.57452) <PHM(0.683)	NBRM(0.151) <PHM(0.288)	PHM
5	NBRM	NBHM	AIC	NBRM(108393)<NBHM(111033)	NBRM(240502)<NBHM(249831)	NBRM
			BIC	NBRM(108503)<NBHM(111245)	NBRM(240622)<NBHM(250063)	NBRM
			McFadden RSQ	NBRM(0.576) <NBHM(0.586)	NBRM(0.151) <NBHM(0.161)	NBHM

Table 3 shows that at stage 1, ZIP outperformed PRM in terms of the Vuong's test, AIC and BIC but the McFadden's RSQ for PRM was slightly higher than that of ZIP for all the sample sizes that are reported in this table. ZIP was selected as the best alternative model to PRM and was compared to NBRM at stage 2 where AIC, BIC and Vuong's test favoured NBRM but the McFadden's RSQ was slightly better than ZIP. NBRM is therefore selected over ZIP based on the three comparison criteria and is compared to ZINB in Stage 3. However, for all sample sizes of at least 60%, ZINB failed to converge or in other words did not reach second order optimality hence there are no parameters reported for ZINB in Table 3. Failure of convergence for ZINB is a clear indication that this model's performance weakens with increased sample sizes. Stage 4 compares NBRM and PHM where the former outperformed the latter in terms of AIC, BIC and Vuong's test but PHM had higher values of McFadden's RSQ. As such, NBRM was compared to NBHM in the last stage and all three comparison criteria favoured NBRM. NBRM is therefore selected as the most effective model for fitting the Duration of Marriage for sample sizes of 60% to 100% and is compared with NBRM for the 10% to 50% sample sizes in the between sample comparison phase.

Between-sample comparison phase: This section compares the models that were chosen under each sample size (see Tables 2 and 3) using the McFadden's RSQ, MSE and MAD and aids in selecting the best model form the ten.AIC and BIC are not used in the between-sample comparison because they are theoretically known to increase as the sample size increases hence they will bias the results when comparing models across different sample sizes. The Vuong's test is also not used for the between-sample comparison because it gave some inconclusive results for some models in the within-sample comparison phase.

Table 4: Comparison of best models the ten samples under study

Sample size		Selected model within a sample	Mc Fadden R^2	MSE	MAD
10%	4392	NBHM	0.59	33.961	3.925
20%	8783	NBHM	0.584	37.604	3.985
30%	13175	NBHM	0.587	36.945	4.03
40%	17566	NBHM	0.554	37.161	4.09
50%	21958	NBHM	0.549	43.761	5.101
60%	25107	NBRM	0.577	35.991	4.031
70%	29311	NBRM	0.575	36.886	4.049
80%	33478	NBRM	0.575	36.971	4.059
90%	37667	NBRM	0.575	36.584	4.053
100%	41881	NBRM	0.151	46.833	5.083

Table 4 shows that for NBHM (10% to 50% sample sizes), the McFadden RSQ generally decreases with an increase in sample size whereas the MSE and MAD increases as the sample size increases. Similar results are observable for NBRM (60% to 100%). This implies that count data models (NBHM and NBRM) generally tend to have smaller McFadden's RSQ values and bigger error margins (MSE and MAD) as the sample size increases. Theoretically, an effective model should minimise the error (MSE and MAD) and maximise the amount of variation explained by the model (McFadden's RSQ) hence, the results in Table 5 imply that NBRM and NBHM become less effective as the sample size increases. Table 5 shows that NBHM for the 10% sample size generally has a better McFadden's RSQ and reduces the error rate much better than other proposed models. As such, NBHM for the 10% sample size from Table 5 is selected as the most effective count data model that can best model the Duration of Marriage as compared to other competing models.

Table 5: Likelihood Chi-Square test results for the selected model (NBHM for the 10% sample size)

Model	Log-Likelihood for the null (Intercept Only) Model	Log-likelihood for the full model	Df	Likelihood Ratio Chi-Square	p-value
10% NBHM	-14900.766	- 6117.149	12	17567.235	<0.0001

Table 6: Parameter estimates of the most effective model

Parameter	Variable	Estimate	Standard Error	DF	t −Value	$Pr > \|t\|$
a11	No of Children	-0.292	0.081	3446	-3.59	0.0003
a0	Logit Intercept	-2.465	0.09	3446	-27.26	<.0001
b0	Log-linear Intercept	0.291	0.06	3446	4.81	<.0001
b3	Male Status	-0.096	0.03	3446	-3.18	0.0015
b4	Female Status	-0.104	0.032	3446	-3.30	0.0010
b5	Male No Times Married	-0.195	0.044	3446	-4.45	<.0001
b6	Female No Times Married	-0.268	0.047	3446	-5.65	<.0001
b7	Male Age	0.2	0.017	3446	11.65	<.0001
b8	Female Age	0.448	0.018	3446	25.19	<.0001
b10	Marriage Type	-0.05	0.01	3446	-5.24	<.0001
b11	No of Children	0.161	0.009	3446	17.30	<.0001
b12	Couple Race	0.087	0.007	3446	11.93	<.0001
v	Dispersion Parameter	7.679	0.353	3446	21.78	<.0001

Determining the significance of the overall model and individual parameters: Table 5 shows that for NBHM of the 10% sample size, the log-likelihood for the full model is -6116.8610 and is -14932.5215 for the null model. The chi-squared value is $2 * \left(-6117.1487-(-14900.7661)\right) = 17567.2348$. Since there are twelve predictor variables in the full model, the degree of freedom for the chi-squared test is 12 yielding a $p - value < .0001$. Thus NBHM for the 10% sample size is confirmed to be statistically significant at 5% level of significance. Table 7 shows all the parameter estimates of NBHM for the 10% sample size. Table 6 shows the parameter estimates for both the log-linear and logit parts of the preferred NBHM. It is worth noting that the log-linear part models the Duration of Marriage (in full years) whereas the log-linear part models the zero-inflation and deviations from equi-dispersion (hurdle part). As such, b1 (Male Occupation), b2 (Female Occupation) and b9 (Solemnisation) are insignificant (at 5% level of significance) in predicting the Duration of Marriage whereas only b11 (number of children) is significant in explaining the hurdle constituent of the preferred NBHM. Figure 1 compares the actual frequencies for the Duration of Marriage to the frequencies estimated using the preferred model as part of the evaluation of the model.

Figure 1: Actual versus NBHM estimated frequencies for the 10% sample size

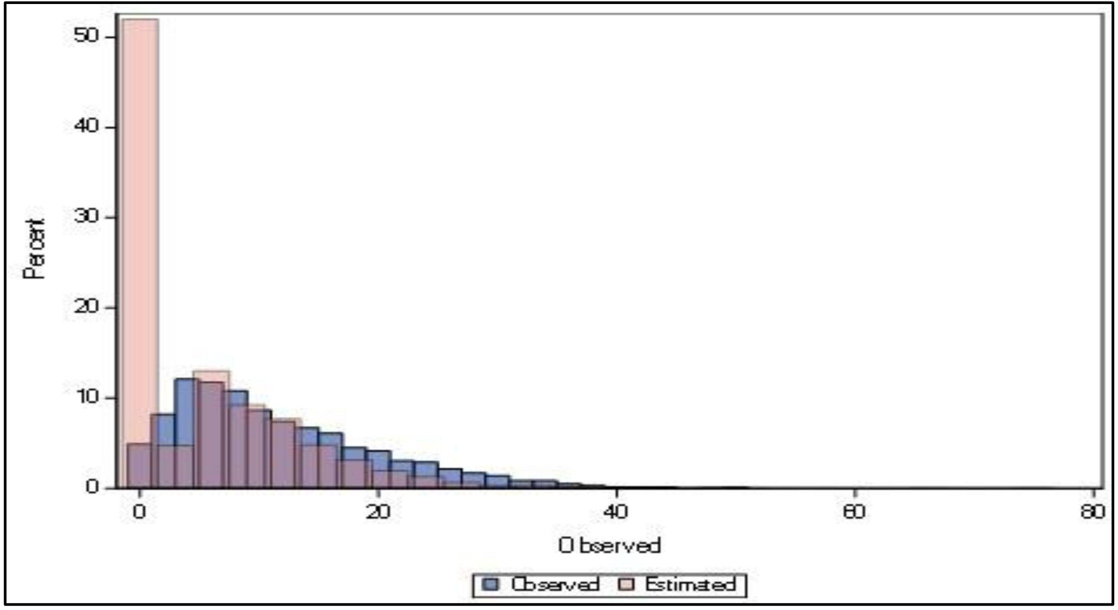

Figure 1 shows that NBHM for the 10% sample size over-estimates the frequency of zero counts, but the estimates of other frequencies are closer to the observed frequencies of Duration of Marriage even though the predicted values are slightly lower than the actual in general.

5. Discussion and Conclusion

This paper generally intended to explore the effect of sample size on the efficiency of count data models. The focus of the study was motivated by the lack of literature about the effect of sample size variations on the performance of popular count data models. Another motivation for this paper is the common practice of having the minimum sample size recommendations as a way of improving the efficiency of multivariate techniques. The study compared the efficiency of PRM, NBRM, ZIP, ZINB, PHM and NBHM under ten sample sizes (4392, 8783, 13175, 17566, 21958, 25107, 29311, 33478, 37667 and 41881) using AIC, BIC, McFadden's RSQ, Vuong's test, MSE and MAD. Empirical findings relative to the within-sample comparison revealed that for sample sizes of 10% (4392) to 50% (21958), NBHM outperformed all models whereas NBRM was favoured by most comparison criteria for sample sizes of at least 60% (25107). The between-sample comparison revealed that generally, the preferred models from the within-sample comparison (NBHM for at most 50% sample size and NBRM for at least sample size) become less effective as the sample size increases. ZINB did not converge when the sample size is at least 50%. The problem of the non-convergence of ZINB is also noted in the study by Famoye and Singh (2006) who also discussed that Lambert (1992) encountered the same challenge. As such, one may remark that ZINB has a disadvantage of not converging especially as the sample size becomes large. NBRM for the smallest sample size under study was selected as the most effective model for fitting the Duration of Marriage and was found to be significant in overall.

Recommendations: Forthcoming research may benefit from applying NBHM which is reported as a better performing model compared to the other five commonly used PRM, NBRM, ZIP, ZINB and PHM. ZINB is the worst performing count data model from the six and as such we suggest that more research should be conducted in order to improve the efficiency of the said model with more focus on its convergence especially in large datasets. The findings of this study revealed that generally, the efficiency of count data models decreases as the sample size increases. As such, sample size has an effect on the efficiency of count data models and imminent research may consider varying sample sizes when applying such models as a way of improving the model selection process. The use of numerous model selection criteria when selecting the optimal count data model may benefit the multivariate analysis by reducing selection bias as opposed using only a few model selection criteria.

This study focused on the six commonly used count data models hence other studies may consider many more count data models such as the Bayesian quintile regression model (Fuzi et al., 2016), the Multivariate Poisson lognormal (MVPLN) (Xiao, Zhang & Ji, 2015) and the Negative binomial-Lindley (NB-L) (Zamani & Ismail, 2010) as alternative models for modelling count outcome data. Future studies may consider the use of other SAS procedures such as GENMOD, GLMIXED, FMM and Macros which can derive count data models as alternatives to NLMIXED which is used in this study. The use of these procedures and other statistical packages such as R and STATA may ease the complexity of deriving the models and probably address issues of non-convergence of ZINB and the computation of Vuong's test statistic for comparing NBHM and other models that were explored in this study. A comparison of the results of the said SAS procedures or statistical packages may help minimise the bias when selecting the optimal count data model.

References

Bajpai, N. (2009). Business statistics. Pearson Education India.

Berk, K. N. & Carey, P. M. (2009). *Data Analysis with Microsoft Excel: Updated for Office 2007*. Cengage Learning.

Burger, M., Van Oort, F. & Linders, G. J. (2009). On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. *Spatial Economic Analysis*, 4(2), 167-190.

Cameron, A. C. & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

Cox, F. & Demmitt, K. (2013). *Human intimacy: Marriage, the family, and its meaning*. Nelson Education.

Famoye, F. & Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4(1), 117-130.

Fuzi, M. F. M., Jemain, A. A. & Ismail, N. (2016). Bayesian quintile regression model for claim count data. *Insurance: Mathematics and Economics*, 66, 124-137.

Hilbe, J. M. (2014). Modelling count data (pp. 836-839). Springer Berlin Heidelberg.

Holman, T. B. (2006). Premarital prediction of marital quality or breakup: Research, theory, and practice. Springer Science & Business Media.

INC, S. I. (2010). SAS/STAT® 9.22 User's Guide.

INC, S.I. (2012). SAS/ETS 12.1 User's Guide.

Little, T. D. (2013). *The Oxford handbook of quantitative methods, volume 1: Foundations*. Oxford University Press.

Liu, W. & Cela, J. (2008). Count data models in SAS. In *SAS Global Forum*, 317, 1-12.

Matignon, R. (2007). Data mining using SAS enterprise miner (Vol. 638). John Wiley & Sons.

Mei-Chen, H., Pavlicova, M. & Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5), 367-375.

Merkle, E. C. & Smithson, M. (2013). Generalized linear models for categorical and continuous limited dependent variables. CRC Press.

Morel, J. G. & Neerchal, N. (2012). Over dispersion models in SAS. SAS Institute.

Ozmen, I. & Famoye, F. (2007). Count regression models with an application to zoological data containing structural zeros. *Journal of Data Science*, 5(4), 491-502.

Park, B. J., Lord, D. & Hart, J. D. (2010). Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. *Accident Analysis & Prevention*, 42(2), 741-749.

Peng, J., Lyu, T., Shi, J., Nagaraja, H. N. & Xiang, H. (2014). Models for injury count data in the US National Health Interview Survey. *Journal of Scientific Research and Reports*, 3(17), 2286-2302.

Reis, H. T. & Sprecher, S. (2009). *Encyclopaedia of Human Relationships: Vol. 1*. Sage.

Rose, C. E., Martin, S. W., Wannemuehler, K. A. & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modelling vaccine adverse event count data. *Journal of biopharmaceutical statistics*, 16(4), 463-481.

Statistics South Africa. (2014). Marriages and divorces, 2012: Metadata/Statistics South Africa. Pretoria: Statistics South Africa.

Stroup, W. W. (2012). Generalized linear mixed models: modern concepts, methods and applications. CRC press.

Tang, W., He, H. & Tu, X. M. (2012). Applied categorical and count data analysis. CRC Press.

Tilanus, E. W. (2008). SET, MERGE and beyond, Proceedings of SAS Global 2008 Conference. Cary, NC: SAS Institute Inc. Paper 167-2008.

Vach, W. (2012). Regression models as a tool in medical research. CRC Press.

Ver Hoef, J. M. & Boveng, P. L. (2007). Quasi-poisson vs. Negative binomial regression: how should we model over dispersed count data? *Ecology*, 88(11), 2766-2772.

Wang, J., Xie, H. & Fisher, J. F. (2011). *Multilevel models: applications using SAS®*. Walter de Gruyter.

Washington, S. P., Karlaftis, M. G. & Mannering, F. (2010). Statistical and econometric methods for transportation data analysis. CRC press.

Whitehead, J., Haab, T. & Huang, J. C. (2012). *Preference data for environmental valuation: combining revealed and stated approaches* (Vol. 31). Routledge.

Xiao, Y., Zhang, X. & Ji, P. (2015). Modelling forest fire occurrences using count-data mixed models in qiannan autonomous prefecture of Guizhou Province in China. *PloS one*, 10(3), e0120621.

Yip, K. C. & Yau, K. K. (2005). On modelling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36(2), 153-163.

Zamani, H. & Ismail, N. (2010). Negative binomial-Lindley distribution and its application. *Journal of Mathematics and Statistics*, 6(1), 4-9.

Zeileis, A., Kleiber, C. & Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software*, 27(8), 1-25.

Appendix 1

Equations for model comparison criteria adopted in this study

1. Likelihood Ratio Test statistic (Merkle and Smithson, 2013)

$$G^2 = -2\left(\mathcal{L}\left(\widehat{\boldsymbol{\theta}}^V|\boldsymbol{y},\boldsymbol{X}\right) - \mathcal{L}\left(\widehat{\boldsymbol{\theta}}^T|\boldsymbol{y},\boldsymbol{X}\right)\right)$$

where $\widehat{\boldsymbol{\theta}}^V$ and $\widehat{\boldsymbol{\theta}}^T$ are maximum likelihood estimates of model V and T respectively. The variables $\boldsymbol{y}$ and $\boldsymbol{X}$ denote the Duration of Marriage and its associated predictor variables respectively.

2. Vuong's test statistic (Little, 2013)

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m}_i)^2}},$$

where m_i for each subject i is calculated as:$m_i = \log\left(\frac{P_1(Y_i|X_i)}{P_2(Y_i|X_i)}\right)$.

3. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Hilbe, 2014).

$AIC = -2\mathcal{L} + 2k$, where$\mathcal{L}$ the log-likelihood function and k is the number of parameters in the model.

$BIC = -2\log\mathcal{L} + k\log n$, where n is the sample size.

4. McFadden *RSQ* (Karlaftis *et al.*, 2010).

$RSQ = 1 - \frac{\mathcal{L}(\boldsymbol{\beta})}{\mathcal{L}(\mathbf{0})}$, where $\mathcal{L}(\boldsymbol{\beta})$ and $\mathcal{L}(\mathbf{0})$ denote the log-likelihood at convergence with the parameter vector $\boldsymbol{\beta}$and the initial log-likelihood with all parameters set to zero respectively.

5. Mean Square Error (MSE) (Wegner, 2007)

$MSE = \frac{\sum(Y_{Ai} - Y_{Pi})^2}{n}$, where Y_{Ai} is the *i*th actual value of Duration of Marriage and Y_{Pi} is the *i*th predicted value of Duration of Marriage.

6. Mean Absolute Deviation (MAD) (Bajpai, 2009)

$MAD = \frac{\sum_{i=1}^{n}|x - \bar{x}|}{n}$,where x is the estimated Duration of Marriage,$\bar{x}$ is the mean DurationOfMarriage and n is the sample size.

7. Wald test statistic Merkle and Smithson (2013)

$\omega = \frac{\hat{\theta} - \theta_0}{\mathrm{var}\,(\hat{\theta})}$, where $\hat{\theta}$ is the maximum likelihood parameter estimate and θ_0 is the hypothesised value.