An Experimental Analysis of the Effect of Sample Size on the Efficiency of Count Data Models

T.V MONTSHIWA

22297812

A thesis submitted in fulfillment of the requirements for the degree *Doctor of Philosophy in Statistics* at the Mafikeng Campus of the North-West University

Supervisor: Prof N.D MOROKE

APRIL 2017

Declaration

I, Volition Tlhalitshi Montshiwa, declare that this thesis is my own work. It is submitted as a fulfilment of the requirements of the degree of Doctor of Philosophy in Statistics at the North West University. It has not been submitted before for any degree or examination in any other university.

.....	.....
Volition Tlhalitshi Montshiwa	Date

Acknowledgements

My heartfelt acknowledgements go to my supervisor Prof N.D Moroke for her constructive inputs, academic support and encouragement throughout the compilation of this thesis. I would also like to thank DataFirst for availing and granting me an opportunity to use their data. I appreciate the North West University for funding my research and providing me with the necessary resources for the completion of this thesis.

Table of contents

List of Acronyms

AIC	Akaike information criterion
AVA	Anthrax vaccine absorbed
AVEerror	Average Error
BIC	Bayesian information criterion
COM-P	Conway-Maxwell Poisson
DIC	Deviance information criterion
DMFT	Decay-missing-filled tooth
FMNB-2	Two-component finite mixture of negative Binomial regression models
GLM	Generalised Linear Models
LAD	Least absolute deviation
LM	Langrage multiplier
LRT	Likelihood ratio test
MAD	Mean absolute deviation
MSE	Mean square error
MSR	Mean squared residual
MVPLN	Multivariate Poisson lognormal
MZIGP	Multilevel zero-inflated generalised Poisson.
MZINB	Multilevel zero-inflated negative Binomial
MZIP	Multilevel zero-inflated Poisson
NBHM	Negative Binomial hurdle model
NB-L	Negative Binomial-Lindley
NBRM	Negative Binomial regression model
PHM	Poisson hurdle model
PLN	Poisson lognormal
PRM	Poisson regression model
QPRM	Quasi-Poisson regression model
RMSE	Root Mean Square error
SAS	Statistical Analyst Software
SITA	State Information Technology Agency

UPB	Unwanted pursuit behaviour
USOs	Unprotected sexual occasions
VIF	Variance inflation factor
ZIDP	Zero-inflated double Poisson
ZIGP	Zero-inflated generalized Poisson
ZINB	Zero-inflated negative Binomial
ZIP	Zero-inflated Poisson

List of Tables

List of figures

Abstract

Many multivariate analysts are of the view that bigger sample sizes yield very efficient models. However, this claim has not been verified for count data models. This study embarked on an experimental analysis of the effect of sample size on the efficiency of the 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 (NBH(M). The study comprised two parts (Part A and Part B). The data used in Part A were sourced from Data First and were collected by Statistics South Africa through the Marriages and Divorces database.

In Part A, the six models were applied to ten random samples selected from the Marriages and Divorces dataset. The sample sizes ranged from 4392 to 43916 and differed by 10%. Part B applied the six models to five simulated datasets with sizes ranging from 50 000 to 1000 000. The models were compared using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Vuong's test, McFadden RSQ, Mean Square Error (MSE) and Mean Absolute Deviation (MAD). The results from Part A revealed that generally, the Negative Binomial-based models outperformed Poisson-based models. However, the results from Part A did not show the effect of sample size variations on the efficiency of the models because there was no consistency in the change in the values of model comparison criteria as the sample size increased. The results from Part B were inconclusive, hence were not meaningful.

Chapter 1

Introduction

1.1 Background and motivation

Count data is defined by Hilbe (2014) to mean observations that only take non-negative integers theoretically ranging from zero to infinity. The author adds that in practice, the upper bound of such data is often limited to the maximum value of the variable being modelled. Hilbe (*ibid*) defines a count variable as a list or array of count data of which in a statistical model the response variable is understood to be random. In other words, the independent values of a count variable in a model can be different at any given time.

Count data are encountered in many studies across disciplines such as population studies and health sciences, to mention a few. An example of count data may be the number of workers who died in mining accidents in South Africa. The number of workers is count data because it can only take positive integers. Tang *et al*. (2012) explain that methods for continuous data such as the classical linear regression model cannot be applied to count responses because count data are unbounded. Another reason why count data may not be analysed with popular prediction models for continuous data is that these type of data often have many zeros, which may be due to the subject not responding to some questions or the respondent not possessing the attribute that the researcher is attempting to measure. Vach (2012) explains that in cases where there are many zeros in count data, the expectation may be negative for some covariates.

Poisson regression model (PRM) is used as the basis for modelling count responses on the assumption that the conditional mean of the outcome variable is equal to the conditional variance (equi-dispersion) holds (Vach, 2012). However, SAS-Institute (2012), Tang *et al.* (2012) and Vach (*ibid*) explain that as much as this is a naturally occurring basic property of a

Poisson distribution, it is not always true in real life data sets. Maumbe and Okello (2013) highlight that count data may exhibit under- or over- dispersion where the latter leads to a larger variance of the coefficient estimates than the expected mean. The authors add that under- or over- dispersion yields inefficient, potentially biased parameter estimates and small standard errors. As such, SAS-Institute (*ibid*) further explains the negative Binomial regression method (NBRM) as an extension of PRM in situations where the variance is significantly bigger than the conditional mean. To elaborate further, Vach (*ibid*) elaborates that the essential assumption of NBRM is that the variances are not equal, but proportional to the mean. According to SAS-Institute (*ibid*), a limitation of both the PRM and NBRM occurs when the number of zeros in the sample exceeds the number of zeros predicted by these models. Wang *et al.* (2011) explain that in the context of count data, the zeros referred to by the previous sentence are termed "excess zeros" or "extra zeros" and the authors explain that the zeros are only structural. The term "structural" means that extra zeros were never observed.

Little (2013) elaborates that excess zeros often occur because values in the sample are from different groups. The author adds that some zeros may come from a group that has a probability of displaying the behaviour of interest, hence always responds with a "0". To elaborate the scenario of excess zeros, consider the previously mentioned example of the study that focuses on fatalities of miners in South Africa. Data that includes the employees such as receptionists in a mining industry will exhibit excess zeros because such employees are not exposed to mining accidents. In that case, one must screen out employees causing structural zeros to remain with the employees who are exposed to the risk of mining accidents.

Little (2013) explains that this screening practice is an issue of study design, and data should be collected to precisely separate structural from non-structural zeros. The author cautions that, excess zeros may cause distorted estimates of the mean and variance of the outcome. As such,

there is a need for count data methods to address the problem of excess zeros. SAS-Institute (2012) adds that the zero-inflated Poisson (ZIP) regression must be used if count data have extra zeros and the assumption of equi-dispersion is not violated. On the other hand, the author explains that zero-inflated negative Binomial (ZINB) regression must be used if the conditional mean and variance are unequal and count data have extra zeros.

Other challenges that may arise in count data modelling are under-dispersion and zero-deflation, but they seldom occur in practice (Ozmen and Famoye, 2007; Morel and Neerchal, 2012). 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 *et al.* (2006). Ozmen and Famoye (2007) and Agresti (2015) explain that the difference between the zero-inflated and the hurdle models is that the latter were formulated to handle zero-deflation as well. In addition, Morel and Neerchal (2012) explain that hurdle models can model both under-dispersion and over-dispersion. Table 1.1 summarises the count data models that have been discussed in this study.

Table 1.1 Popular Count data models

Category	Model	Designed for data that are:
Basic	PRM	Equi-dispersed, Not zero-inflated
	NBRM	Over-dispersed, Not zero-inflated
Zero-Inflated	ZIP	Eqi-dispersed, Zero-inflated
	ZINB	Over-dispersed, Zero-inflated
Hurdle	PHM	Under-/Over-dispersed, zero-inflated/deflated
	NBHM	Under-/Over-dispersed, zero-inflated/deflated

Table 1.1 shows PRM and its extensions which were founded in an attempt to address the shortcomings of this basic model. None of the models discussed hereto are superior as there is an on-going criticism which leads to iterative formulation of new models. However, there is no convergence on whether or not a certain model can address all the problems encountered in count data. On the contrary, the present study acknowledges that the efficiency of count data models is mainly affected by poor data quality (excess zeros) and violations of distributional assumptions (under- or over- dispersion, for instance), hence an effort should be made to improve on data quality than just re-parameterisation of count data models.

One known way of improving the efficiency of multivariate prediction methods such as multiple regression models is through an increase in the sample size. As such, this study generally seeks to revisit the fundamental PRM and its extensions (NBRM, ZIP, ZINB, PHM and NBHM) in an attempt to find out whether or not an increase in sample size can improve the efficiency of these models relative to under- or over- dispersion and excess zeros. In addition to the lack of convergence on determining the ideal count data model, this study recognises the seldom use of count data models in the context of marriage data. As such, the study further seeks to apply count data models to marriage data, having in mind the advantage of the inclusion of both categorical and continuous variables in such models.

The remaining sections are to discuss the following: Section 1.2 provides the problem statement followed by aim and the objectives of the study in Section 1.3. Section 1.4 presents the rationale of this study. The significance of this study is outlined in Section 1.5 and Section 1.6 presents the assumptions. The structure of this study is presented in Section 1.7 and the summary of this chapter is presented in Section 1.8.

1.2 Problem Statement

The most important part of multivariate analysis is selecting the right technique for a particular type of data set in order to obtain more realistic, valid and reliable results. At present, the fundamental differentiator of methods used in modelling count data is the distributional assumptions of the mean and variance, the presence of excess zeros and the theoretical knowledge of the data set. There are disagreeing conclusions in literature in terms of which count data model is the best. The performance of the extensions of PRM which are meant to address its limitations such as its non-applicability to under-/ over-dispersed data and excess zeros is continuously questioned and new models are iteratively established in order to improve such extensions of PRM.

The re-parameterisation of count data models has not yet led to a common decision on the ideal model for handling both under- and over- dispersion and effects of excess zeros. Therefore, there is need for a study that can explore ways of improving the efficiency of count data models without re-parameterisation of existing models. This study proposes sample size as part of considerations in analysing count data. This proposition arises from the common practice of sample size recommendations which often form part of criteria for improving the efficiency of multivariate techniques.

In most cases, the bigger the sample size the more robust the techniques. Examples of such techniques include factor analysis and multiple regression analysis, to mention a few. As such,

this study largely seeks to understand whether or not the sample size variations can improve the efficiency of count data models relative to under- or over- dispersion and excess zeros without further iterative re-parameterisation of the known models. In addition, this study seeks to illustrate the applicability of count data models to marriage data as it proposes that the DurationOfMarriage is count data and should be treated as such. Count data models discussed hereto are seldom applied to divorce data despite their flexibility in using both categorical (common in divorce data) and continuous variables as predictors. In summary, this study is set to answer the following question: Does sample size affect the efficiency of count data models?

1.3. Aim and objectives of the study

The aim of this study is to embark on an experimental analysis with the intent to explore the effect of sample size on the efficiency of count data models when the data exhibit under- or over- dispersion as well as zero-inflation/deflation. This will provide a platform for deciding if there is a need for a specific minimum sample size recommended for count data models to be efficient.

This study is set to address the following objectives:

1.3.1 To conduct an experimental analysis by exploring the efficiency of count data models under different sample sizes from original and simulated data.

1.3.2 To use the theoretical framework of count data to assist in building the six count data models and predicting the DurationOfMarriage.

1.3.3 To identify the key predictors of the DurationOfMarriage from the available predictor variables.

1.3.4 To identify the most efficient count data model for predicting the DurationOfMarriage.

1.3.5 To use the findings in formulating suggestions for future studies and parties involved in marriages and divorces.

1.4. Rationale of the study

This study is conducted to assess the impact of an increase in sample size on the efficiency of the most commonly count data models. The study is conducted with the concern that even though sample size requirements form part of the assumptions of many multivariate methods, its impact on the efficiency of count data models has never been considered. It is therefore important to explore the effect of sample size variations on the efficiency of count data models, more especially when the target variable is zero-inflated and/or over-dispersed. This long overdue study is therefore an important guide to other researchers who are interested in either implementing or comparing count data models.

1.5. The significance of the study

The findings of this study will serve as reference for researchers interested in determining the efficiency of count data models on similar data used in this study. Furthermore, scholars who are interested in employing the ideal model for marriage data, more specifically when the intention is to predict the duration of a marriage will, benefit from the findings of this study. This study will assist researchers when deciding on the minimum sample size necessary for count data model to be more efficient when the data exhibits under- or over- dispersion. The findings of the study will also assist researchers in comparing the results of count models from actual and simulated data sets.

The findings may also be of value to marriage councillors or married couples as different models will highlight the key predictors of the DurationOfMarriage accordingly. This could provide guidance to even people who are planning to marry, not to start on the wrong footing.

This knowledge about the effect of predictors of the DurationOfMarriage (the socio-demographic and economic characteristics of the couple) will inform the couples about aspects needing improvement in an endeavour to a lengthy or a lifetime marriage. By so doing, the couples can work on the contributing factors so as to minimise the probability of a divorce.

This study will also serve as an eye opener to data capturers and analysts that DurationOfMarriage recorded as full years is count data, and should be treated as such. Guidance may also be provided on the optimal use of the models when predicting the DurationOfMarriage. The use of marriage data in count data models is also significant when utilising categorical variables as predictor variables, which is not allowed in the popularly used linear regression analysis. Policymakers may use the predictive model to manage the possibility of risks associated with divorce. For example, insurance companies may predict the DurationOfMarriage and decide on whether a certain couple should qualify for a certain policy.

1.6 Assumptions of the study

This study assumes that the data used is from a homogeneous sample because it is about the marriages and divorces that occurred in South Africa for the period 2010 and 2011. It is also assumed that the data were collected using the same instruments for the two time periods. The study also assumes that DurationOfMarriage is count data because it is counted as non-negative integers only (in full years, not decimals or fractions). The models used in this study are assumed to be suitable for count data as informed by literature and they share similar characteristics because they all belong to the Generalised Linear Models (GLMs) family.

1.7 Structure of this study

The remaining chapters of this study are to discuss the following: a review of literature is given in Chapter 2; Chapter 3 discusses the statistical methodology implemented in this study; the

data analysis results are presented in Chapter 4; and the findings, discussion as well as the recommendations for further research are provided in Chapter 5. The reference list for studies that were interrogated in this study is provided after Chapter 5. Appendix 1 presents the general SAS Codes used in this study followed by Appendix 2 which presents the first 20[1] observations of the merged data set used in this study.

1.8 Chapter summary

This chapter presented the background and motivation of the study. The six proposed count data models which are PRM, NBRM, ZIP, ZINB, PHM and NBHM were explained in detail through highlighting the reasons why each model was developed and the shortcomings of each model were also discussed. The chapter also highlighted the reasons that motivated the undertaking of the study, the aim and objectives that are set to address the identified problem and the proposed methodology for achieving the objectives of this study were discussed. The significance, contribution, limitations of this study were outlined in this chapter as well. The next chapter reviews literature on the subject.

[1] The merged data set is very large, hence only the first 20 observations are shown. The complete data sets used in this study may be made available on request.

Chapter 2

Literature review

2.1 Introduction

This chapter deliberates on literature around the predictors of a successful marriage in section 2.2 and comparative studies which focused on the methods used for modelling count data in Section 2.3. However, the techniques used in determining the predictors of marriage success are not of interest to this study, the interest is mainly on the incidence of variables that were recommended by many studies in literature. The general reason for reviewing literature on the key predictors of a successful marriage is to support the use of the available socio-economic and demographic variables in the data set used in this study to predict the DurationOfMarriage. Previous studies around the comparison of PRM, NBRM, ZIP, ZINB, PHM, NBHM and other count data models are reviewed with the intent of identifying similarities and differences between them. Reviewing literature on previous comparative studies around count data models also aids in highlighting that such studies did not focus on the effect of sample size on the efficiency of the proposed count data models used in this study.

2.2 Predictors of marriage

Cox and Demmitt (2013) discuss that literature has identified that background factors such as age at marriage, education, economic class, and race have some influence in the success of a marriage. The authors also added that factors such as occupation, friends and religion have some marital predictive value. Reis and Sprecher (2009) discuss that researchers have found that race, level of education and gender are significant predictors of the risk of divorce over time. In addition, Reis and Sprecher (*ibid*) point out that factors such as passion, liking, trust and emotional distress (clustered as feelings) are efficient in predicting relationship outcomes.

Other predictors of divorce as identified by Reis and Sprecher (*ibid*) include personal traits, communication and support as well as physical aggression. Holman (2006) identifies the predictors of quality of marriage as social network support (examples being support from family, co-workers and friends) and socio-cultural context (with age at marriage, education, income, occupation, race religion and gender being highlighted).

It is evident from literature that, although they are not the sole predictors of marriage success or divorce, socio-economic and demographic attributes of the couples have been identified as having significant contribution to the success of a marriage (Holman, 2006, Reis and Sprecher, 2009 and Cox and Demmitt, 2013). As such, this study uses socio-economic variables such as male occupation and demographics such as female age (based on availability) to fit count data models for predicting the DurationOfMarriages in South Africa. It is worth highlighting that the prediction of DurationOfMarriage is a secondary aim of this study, but the main focus is on comparing the proposed count data models. The next section gives a thorough review of previous studies which focused on comparing count data models of interest.

2.3 Empirical Literature

Burger *et al.* (2009) conducted a study to compare the performance of Poisson-based and Binomial-based models. The authors used geographical distance, contiguity, common language, common history, free trade agreement, institutional distance and sectoral complementarities to predict the Yearly average volume of trade for the period 1996 to 2000. Burger *et al.* (2009) used multiple linear regression coefficients and two measures of goodness-of-fit, namely: residuals and Stavins and Jaffe goodness-of-fit statistic as the basis for comparing the methods under study. The Akaike Information Criterion (AIC) and Bayesian

Information Criterion (BIC) were also used to compare the efficiency of the Poisson-based and Binomial-based models.

The results of the study by Burger *et al.* (2009) revealed that, compared to the PRM estimator, the regression coefficients estimated by ZIP were similar, while the regression coefficients estimated by NBRM and ZINB differed substantially from those of PRM. The authors found that PRM and ZIP perform relatively well, as the estimated volume of trade does not deviate much from the observed volume of trade for either small or large trade flows. The Stavins and Jaffe goodness-of-fit statistics used in the study under review were based on the Theil inequality coefficient (Theil's U). Burger et al. (*ibid*) found that the value of the Stavins and Jaffe goodness-of-fit statistic obtained from the PRM and ZIP models is significantly higher than the values obtained from the NBRM and ZINB models. Another criterion to evaluate the performance of the count data models under study was a comparison of the expected probabilities to the observed probabilities for each model.

Burger *et al.* (2009) explain that the points above the x-axis represent an over-prediction of the probability of observing that volume of trade, while the points below the x-axis represent an under-prediction. Burger et al. (*ibid*) deduced from their results of the study that ZINB performs the best, followed by NBRM and ZIP, which performed about equally well. In addition to the graphical methods, the authors examined more formal statistics namely, the likelihood ratio and the Vuong tests for over-dispersion. The results of the study by Burger *et al.* (*ibid*) revealed that the likelihood ratio test for over-dispersion and the Vuong test indicated that NBRM is favoured over PRM, ZIP is favoured over PRM, and ZINB is favoured over NBRM, ZIP and PRM.

The authors noticed that neither the Vuong nor the likelihood ratio tests for over-dispersion can be used in comparing NBRM and ZIP. This current study intended to confirm this claim by the

authors. The authors found out that both the AIC and the BIC indicate that the NBRM should be preferred over the ZIP. The authors commented that overall, it can be inferred that ZIP performs the best on average, as rated by both criteria. This is because the authors found that it has a reasonable fit of estimated trade, can include zero flows, and accounts for different types of zero flows, correcting for excess zeros and the over-dispersion that results from that. The current study is different from that of Burger *et al.* (2009) because it also explores the efficiency of hurdle models and it compares count data models under different sample sizes.

Rose *et al.* (2006) conducted a study to compare and contrast several modelling methods for vaccine adverse event count data with over-dispersion. Rose et al. (*ibid*) analysed data from an anthrax vaccine absorbed (AVA) clinical trial study, in which the number of systemic adverse events occurring after each of four injections was collected for each participant. The study assessed the model fit of PRM, NBRM, ZIP, ZINB, Poisson hurdle model (PHM) and Negative Binomial hurdle model (NBHM). Rose *et al.* (*ibid*) clarify that unobservable heterogeneity, which leads to over-dispersion, is likely to be an issue since the AVA trial enrolled participants with significant variation in their socioeconomic and health related factors. In addition, the study gathered that temporal dependency due to multiple injections over time for each participant may be an issue. Furthermore, the authors clarified that there may be excess zeros because it is expected that many participants will not experience any systemic adverse events during the time periods monitored.

Rose *et al.* (2006) implemented the six methods in modelling the count of systemic events occurring after a dose for a participant. The explanatory variables used are treatment (Groups I–V), study centre, gender, race and time. The authors explain that all variables were categorical except for time. Criteria for determining the excess zeros and over-dispersion were chosen based on whether or not the models were nested within others. Rose *et al.* (*ibid*) explain that

the PRM and NBRM models are nested within the ZIP and ZINB, respectively. As such, Rose *et al.* (*ibid*) tested for over-dispersion due to excess zeros using the likelihood ratio and score tests by comparing the PRM and NBRM models to the ZIP and ZINB models respectively. The results showed that the score and likelihood ratio tests both favoured the ZIP and ZINB models ($p < 0.0001$), hence the authors concluded that there was evidence of over-dispersion due to excess zeros.

Rose *et al.* (2006) explain that PRM and ZIP are not nested within PHM, also NBRM and ZINB are not nested within NBHMs. The Vuong statistics for the PHM versus PRM and NBHM versus NBRM in the study by Rose *et al.* (*ibid*) were found to be in favour of the hurdle models. The Vuong statistics for PHM versus ZIP, and NBHM versus ZINB revealed that neither model is favoured. The Vuong statistics revealed that there is significant over-dispersion due to both heterogeneity and excess zeros. This was observed from the Vuong statistics favouring negative Binomial models over Poisson models and zero-inflated/hurdle models favoured over the standard PRM and NBRM. Subsequent to testing for over-dispersion and excess zeros, the authors examined the model fit using AIC and BIC and compared expected probabilities and resulting counts for each model with observed counts. The authors gathered that AIC criterion favours the ZINB and NBHM over all other considered models. However, the authors noticed a little difference between the AIC values of ZINB and NBHM. On the contrary, the results of the study revealed that the BIC criterion favoured the NBRM but the BIC values for ZINB and NBHM were found to be close to the NBRM value.

Rose *et al.* (2006) found that the NBRM predicted the observed frequencies for counts greater than zero adequately, but that there were some unexplained zeros. The number of zero counts predicted by zero-inflated and standard models was found to be close to the observed number of zeros. However, the results showed that the ZINB model and NBHM fit better than the ZIP

and PHM for all other count categories. The generalised Pearson chi-square test indicated that the PRM, ZIP and PHM models exhibited lack of fit p-values < 0.0001. NBRM was found not to exhibit lack of fit but the ZINB and NBHM were found to improve the fit substantially. In addition, the results showed that the predicted mean for each model was close to the empirical expected value, but the variance predicted by the four models varied substantially from the empirical variance estimate. Based on the results of their study, Rose *et al.* (*ibid*) concluded that zero-inflated models or PHM can account for over-dispersion resulting only from excess zeros. However, the authors commented that zero-inflated or hurdle negative Binomial models are the most flexible of the considered models because they can account for over-dispersion from excess zeros and unobserved heterogeneity.

Rose *et al.* (2006) noted that the fitted models revealed that zero-inflated and hurdle models are indistinguishable with respect to goodness of fit measures. However, the authors advised that choosing between the zero-inflated and hurdle models, assuming the PRM and NBRM are inadequate because of excess zeros, should generally be based on study endpoints and goals. To elaborate this statement, the authors explained that if the goal is to develop a prediction model, then it is not important which modelling framework to use (*i.e*. zero-inflated or hurdle), assuming predictions are indistinguishable. However, the authors added that if the goal is inference, then it is important to choose the model that is most appropriate given the study design.

The results from the study by Rose *et al.* (2006) should be used cautiously because the authors disclosed that for the purpose of demonstration. In addition the phrase "Our study illustrates, for our data, the superiority of standard, zero-inflated, and hurdle negative Binomial models over the standard, zero-inflated, and hurdle Poisson models" in the study by Rose *et al.* (2006, 478) implies that the study by Rose *et al.* (*ibid*) is a confirmatory study. However, the present

study is experimental and is generally aimed at enquiring about the effect of sample size variations on the efficiency of count data models.

Hoef and Boveng (2007) compiled a statistical report with the objective of introducing some concepts that may assist ecologists in choosing between a quasi-Poisson regression model (QPRM) and NBRM for over-dispersed seal count data. The target variable was seal counts and exploratory variables were date, time of day and tide. The predictor variables were determined based on theoretical knowledge of ecology. This way of selecting predictors is also applied in the current study such that the exploratory variables are chosen based on theoretical knowledge suggested by literature.

In an attempt to determine the extent to which the use of the two models under discussion affect the fitting of the regression coefficients, the authors examined the graphs of mean-to-variance relationship and the weights as functions of mean. The results of the graphs in the study by Hoef and Boveng (2007) showed that there is a difference in the mean or variance relations of NBRM and QPRM. As such, Hoef and Boveng (*ibid*) concluded that regression coefficients might be fit differently between NBRM and QPRM because fitting these models uses weighted least squares. The authors elaborate that these weights are inversely proportional to the variance. As such, the Hoef and Boveng (*ibid*) emphasized that NBRM and QPRM will weigh the observations differently. In their discussion, the authors deliberated that there is no general way to determine the best model from QPRM and NBRM. However, the authors explained that based on their example, QPRM is a better fit to the overall variance-mean relationship.

Hoef and Boveng (2007) affirm that ultimately, choosing among QPRM, NBRM and other models is a model selection problem. The authors advised that although they have pointed out the shortcomings o likelihood based models, other approaches that do not depend on distributions, such as cross-validation, could be used in addition to diagnostic plots used in

their example. Moreover, the authors emphasized that an important way to choose an appropriate model is based on sound scientific reasoning rather than a data-driven method. The authors commented that the QPRM formulation has an advantage of leaving parameters in a natural, interpretable state and allows standard model diagnostics without loss of efficient fitting algorithms.

The report by Hoef and Boveng (2007) concurs with the study by Rose *et al.* (2006) that the choice of the method to use for modelling count data depends on the theoretical knowledge of the topic at hand. On the other hand, Hoef and Boveng (*ibid*) discourage the use of distribution-based criteria such as likelihood methods, AIC and BIC because such criteria do not always precisely differentiate count data models. However, both Rose *et al.* (2006) and Burger *et al.* (2009) used AIC and BIC. This highlights the need to critically choose the criteria for comparing count data models in the current study. More literature is needed on the comparison criteria of count data models in the current study. The reasoning relative to the mean and variance characteristics of the models studied by Hoef and Boveng (*ibid*) shed the light that these characteristics need to be considered when dealing with the disadvantage of over-dispersion in the standard PRM. As such, this current study ensures that the mean and variance are approximately equal across all ten sample sizes that are proposed for this study in order to reduce bias in capturing the effect of sample size on the efficiency of count data models.

Park *et al.* (2010) conducted a study to investigate the potential bias and the variability in parameter estimates in the finite mixture models using various combinations of sample sizes under varied sample-mean values. The main focus area of the study by Park *et al.* (*ibid*) was the bias associated with the posterior mean and median of dispersion parameters in the two-component finite mixture of negative Binomial regression models (referred to their study as FMNB-2). The authors explain that FMNB-2 models are known to be flexible and allow one

to capture heterogeneity through usually a small number of simple regression models such as PRM or NBRM. Park *et al.* (*ibid*) argue that, from an application-oriented point of view, it is important to know the minimum sample size necessary in order to guarantee the unbiased or bias-reduced estimates of model parameters. This statement by Park *et al.* (*ibid*) supports the main aim of the current study, which is to determine the sample size requirements for count data models. The authors were also interested in the sample mean values. Park *et al.* (*ibid*) note that, within the standard negative Binomial modelling framework, several researchers have found that the dispersion parameter in the NBRM is significantly influenced by not only sample sizes, but sample mean values as well. This current study therefore takes this argument in consideration by ensuring that the sample means are approximately equal across all the sample sizes considered.

Park *et al.* (2010) used a Monte Carlo simulation to generate various sample sizes under different sample-mean values. In addition, the authors investigated two different prior specifications for the dispersion parameters: non-informative and weakly-informative gamma priors. As the authors explain, their interest in these priors is informed by literature that prior specification for the dispersion parameter has a potential influence on the posterior summary statistics. As such, the intention of their study was to compare results from non-informative and weakly-informative prior specifications in terms of the magnitude of the bias introduced by various sample sizes and sample-mean values. Park *et al.* (2010) first designed the simulation scenarios for generating FMNB-2 random variates. The authors explain that the regression parameters, mixing proportions, and dispersion parameters were controlled in order to generate three sample-mean categories namely: high mean ($\bar{y} > 5$), moderate mean ($1 < \bar{y} < 5$) and low mean ($\bar{y} < 1$). The study further explained that, in order to allow for a high level of heterogeneity, the higher-mean component (Component 1) was combined with a lower dispersion parameter and the smaller-mean component (Component 2) was combined

with a higher dispersion parameter. The authors replicated the FMNB-2 random variable generation process 100 times for each category, and then for each of the data sets a Bayesian estimation was carried out using 2500 draws after a burn-in of 2500 draws. It is explained that, at the end of each replication, the posterior summary statistics such as posterior mean, median, standard deviation for each parameter estimate were computed.

Park *et al.* (2010) used the mean square error (MSE) to check the quality of the estimator because as the authors explain, this criterion comprises parameters for both bias and variability. The results of their study were such that for the high sample-mean value scenario, the bias was negligible for the higher-mean component (Component 1) except when the sample size was too small (about $N = 300$) for both priors. However, the study noticed that the bias was more significant in the smaller-mean component (Component 2), but this was particularly true when the posterior mean was used as a summary statistic with the non-informative prior. Moreover, for the non-informative prior case, there was an upward-bias trend for both posterior mean and median in Component 2. The simulation study conducted on the FMNB-2 model showed that the posterior mean using the non-informative prior exhibited a high bias for the dispersion parameter, especially in the smaller-mean value component.

The posterior median instead was found to have much better bias properties than the posterior mean, particularly at small sample sizes and small sample means. However, Park *et al.* (2010) noticed that as the sample size increased significantly for both small to moderate mean value scenarios, the posterior median using the non-informative prior also began to exhibit the upward-bias trend. The authors explain that this is because as the sample size increases, the posterior median is getting closer to the posterior mean, which exhibits the upward-bias. The authors further elaborated that the use of the weakly-informative prior had the advantage of reducing the variability in the estimates for the posterior mean and median, but it tended to

underestimate the true value by pulling the estimates toward its prior mean. The authors also noticed that as the sample-mean value decreases, this tendency was found to be more pronounced.

The current study is similar to the study by Park *et al.* (2010) because it also uses count data models under different sample sizes. However, as opposed to the study by Park *et al.* (*ibid*), the current study compares numerous count data models, but simulated data is only used as an extension of the actual data case. Also, the sample means as well as variances are kept approximately equal across all sample sizes in this study as opposed to the one by Park *et al.* (*ibid*). The reason for keeping the means and variances constant across samples is to minimise bias and to focus exclusively on the effect of sample size on the efficiency of the models without external influences.

Mei-Chen *et al.* (2011) conducted a study to illustrate the differences between PRM, NBRM, ZIP, PHM, ZINB and NBHM as well as to explore how to compare different models. The authors used data from a multisite clinical trial of behavioural interventions to reduce episodes of HIV-risk behaviour. There were 515 subjects in the data. The outcome was the count of unprotected sexual occasions (USOs) with male partner(s) measured at three and six month follow-up points after one of the intervention programmes (Sex Skills Building versus HIV education) was implemented. The predictor variables were the intervention condition, time, count of USOs at starting point and age. Over-dispersion in the PRM was tested by the Lagrange multiplier (LM) statistic. Of all the studies reviewed hereto, the LM statistic is only implemented in the study by Mei-Chen *et al.* (*ibid*), which suggests that the methods for determining the occurrence of over-dispersion in the data are not exhausted in the studies reviewed hereto. Therefore, there is a need for the current study to review more literature on

such methods. For negative Binomial models, the dispersion parameters were tested for difference from zero with t-statistics.

The LRT (for full and nested models), AIC and Vuong statistics (for non-nested models) were used in comparing the goodness of fit between pairs of models. According to the LM statistic as used in comparing pairs of full and nested models namely: NBRM versus PRM, ZINB versus ZIP and NBHM versus PHM, the ZINB model showed superior fit. The PRM was found to be inferior to all other models. The authors remarked that zero-inflated models fit better than their corresponding non-zero-inflated counterparts, which suggests that the best-fitting model needs to account for both over-dispersion and zero-inflation in the observed data. The AIC and Vuong tests revealed that ZIP and ZINB models fit better than PHM and NBHM respectively. Mei-Chen *et al.* (2011) remarked that this suggests that the zero counts were best modelled, not only from structural zeros as in the hurdle models, but as being due to both structural and sampling zeros. The main difference between the study by Mei-Chen *et al.* (*ibid*) and the current study is that the former used a single data set whereas the latter extends the literature by examining the efficiency of count data models under various sample sizes.

Famoye and Singh (2006) conducted a study to develop a zero-inflated generalized Poisson (ZIGP) regression model for modelling over-dispersed with too many zeros. The authors used domestic violence data with 214 cases. The dependent variable, violence, was the number of violent behaviour of barterer towards victim. The predictor variables used in the regression models were level of education, employment status, level of income, having family interaction, belonging to a club, and having drug problem. The authors compared only three count data models namely: ZIP, ZINB and ZIGB. Famoye and Singh (*ibid*) explain that they opted to include the zero-inflated models because the data had too many zeros (observed proportion of zeros was 66.4%). However, the authors explain that ZINB did not converge and they add that

Lambert (1992) also observed this problem in fitting ZINB model to an observed data set. The authors motivate that they therefore had to develop and to apply the ZIGP regression model for modelling over-dispersed data with too many zeros. This implies that the ZIGP was therefore developed to improve the shortcoming of the ZINB relative to convergence.

In order to confirm the statistical significance of the 66.4% observed proportion of zeros, Famoye and Singh (2006) conducted a score test for zero inflation in the PRM. The authors elaborate that the score test aims to check whether the number of zeros is too large for a PRM to adequately fit the data. The results showed a significant score test statistic implying that the data have too many zeros and the PRM model is not an appropriate model. It is elaborated in the study by Famoye and Singh (*ibid*) that the ZIGP reduces to the ZIP when the dispersion parameter $\alpha = 0$. As such, the authors proposed that the null hypothesis of $\alpha = 0$ be tested in order to determine the adequacy of the ZIGP over the ZIP. The Wald statistic was used in this regard. The Wald test rejected the null hypothesis, hence the authors concluded that the ZIP was inappropriate for modelling domestic data and proposed that the ZIGP is ideal. The authors also used the log-likelihood statistic to compare the goodness of fit for the two models and they found that the ZIGP fits the data better than the ZIP model with almost double the value of the likelihood of the latter.

Famoye and Singh (2006) concluded that generally the ZIGP is a good competitor for the ZIP. However, the authors were unable to compare ZINB with ZIGB and ZIP because ZINB failed to converge. Of all the studies reviewed hereto, only the study by Famoye and Singh (*ibid*) has introduced a completely new model, the ZIGP. The ZIGP is not part of the current study because the intention is to focus on the popular and mostly implemented count data models. The current study also diverges from the study by Famoye and Singh (*ibid*) because it focuses on different sample sizes and many other count data models.

Yip and Yau (2005) conducted a study to explore the application of zero-inflated models on insurance claim count data. The authors considered four zero-inflated models namely: ZIP, ZINB, ZIGP and zero-inflated double-Poisson (ZIDP) which were also compared with the PRM and NBRM. The authors used the motor insurance claim frequency data set which is retrieved from the SAS Enterprise Miner database. There were 2812 complete records that were drawn from the SAS data set which comprised 33 variables. The outcome variable was the count of claims made by a policyholder in the policy year and there were 13 predictor variables which included the usage of the car, marital status, residential area, income and gender of policyholders, to mention a few. The authors explain that the variables providing a significant improvement in the Poisson's log-likelihood function at convergence were chosen as predictor variables. As in the study by Famoye and Singh (2006), Yip and Yau (*ibid*) used the score test statistic for zero-inflation in the PRM. The authors concluded from the results of the score test that there was enough evidence of the existence of too many observed zeros.

Yip and Yau (2005) evaluated the performance of the PRM, NBRM, ZIP, ZINB, ZIGP and ZIDP models using the log-likelihood, AIC, BIC and the generalized Pearson chi-square statistic. The authors found that the NBRM and the zero-inflated regression models fit the motor insurance data reasonably well with the ZIDP regression model providing the best fit to the data. One may notice that, among all the studies reviewed hereto, the ZIGB is considered in only two studies namely: Yip and Yau (*ibid*) and Famoye and Singh (2006). The ZIDP is considered for the first time in the study by Yip and Yau (*ibid*). This clarifies the claim made in the current study that count data models are not exhaustive and that no matter how many models are derived, there is still no convergence in the conclusions made by the authors about the best count data model. This trend of disagreeing conclusions in literature supports the need for the current study to explore a different way of improving efficiency of the popular models

(sample size variations) rather than only relying on altering the parameterisation of the existing models.

Potts and Elith (2006) conducted a study to compare the performance of some count data regression methods within a generalised linear model framework namely: PRM, NBRM, QPRM, PHM and ZIP. The outcome variable was the count of rocky outcrops with *L.ralstonii* present in a 400m radius. The authors used variables that were previously found to be significant predictors of the distribution and abundance of rocky outcrop species namely: rainfall and outcrop area. The authors emphasise that the choice of predictor variables was fixed for each model so that differences between the models and their predictions could be attributed to differences in model specification. The same approach is used in this present study in order to measure the impact of sample size on the performance of the models without noise that may arise due to differences in the choice of variables. Potts and Elith (*ibid*) explain that all models of interest were implemented in R. On the other hand, the data in this current study is analysed using SAS based on the author's preference.

Potts and Elith (2006) assessed the presence of over-dispersion by computing the ratio between the mean and variance. The authors found out that the data were over-dispersed because the variance was much greater than the mean. It is worth noting that the authors did not specify how large "much greater than" was, therefore this way of assessing over-dispersion should be used with caution. The use of rules of thumb may be important. On the other hand, this current study uses the LRT and Vuong's test to examine over-dispersion in nested and non-nested models, respectively. Potts and Elith (*ibid*) assessed the possibility of zero-inflation by visually inspecting a histogram of the data. The authors concluded that the data were zero-inflated due to a spike at 0 in the histogram. This current study also implements the histograms in determining whether zero-inflation exists in the data. Pearson correlation coefficient (*r*),

Spearman's rank correlation (ρ), model calibration, Average Error (AVEerror) and Root Mean Square error (RMSE), were used as criteria for comparing the models. The *r* was used to determine how closely the observed and predicted values agree and ρ was used to determine the similarity between the ranks of the observed and predicted values. On the contrary, this current study compares the frequencies of the observed and predicted values as opposed to the *r* and ρ as another way of assessing the dispersion of the estimated values around the actual values. Model calibration was assessed by fitting a simple linear regression between the observed and predicted values which was such that observed = *m* (predicted) + *b*, where the values of *m* and *b* provide information about the degree of bias. Models with smallest amounts of RMSE and AVEerror are preferred.

The study by Potts and Elith (2006) revealed that PHM performed best because both the *r* and ρ showed that predictions and observations were relatively similar in magnitude and similarly ordered. The model calibration for PHM indicated a relatively small but consistent bias and PHM had the smallest RMSE but a slightly large AVEerror relative to other competing models. The authors identified NBRM as the worst performing model with a weak *r*, the model calibration which indicated a strong and inconsistent bias, the highest RMSE and AVEerror values relative to all models tested. Only ρ was found to be as high as that of PHM. Potts and Elith (*ibid*) also noticed that the ZIP model had a lower *r* and ρ despite having the best model calibration of all models tested. The authors explain that these ZIP outcomes are as a result of the amount of error around the predictions being high but the AVEerror were accurate.

Potts and Elith (*ibid*) also found that PRM and QPRM were comparable in performance, mainly because they had the same parameter coefficients. The authors explain that both models had predictions that were relatively dis-similar in value but similar in rank as observed in the values of *r* and ρ, as well as by relatively consistent medium biases in the model calibration

parameters. On the other hand, the authors noted that PRM and QPRM had small RMSE and AVEerror because when averaged across all locations, their mean predictions were relatively accurate. The study by Potts and Elith (*ibid*) differs to most previous studies because it used *r*, ρ, RMSE, AVEerror and model calibration as opposed to the common count data model fit criteria which includes *inter-alia* AIC and BIC.

Fuzi *et al*. (2016) conducted a study to compare the Bayesian quantile regression model to PRM and NBRM using the Malaysian motor insurance claims data. The authors explain that GLMs such as PRM and NBRM are mean regression models in which the mean is explained by a set of covariates that are associated with a link function. However, Fuzi *et al*. (*ibid*) clarify that quantile regression models utilise the least absolute deviation (LAD) instead of the least square error which is usually used in ordinary mean regression models. An advantage of quantile regression as highlighted by the authors is that it does not require any type of distribution hence it does not depend on any property of distribution. In the study by Fuzi *et al.* (*ibid*), risk exposure was modelled using vehicle year, vehicle cc, vehicle make and the location as predictor variables.

Fuzi *et al.* (2016) used the LRT for comparing PRM and NBRM. In order to assess the overall performance of all models under comparison, the authors compared the actual and estimated claims counts. The results of the LRT indicated that the claims counts are over-dispersed hence NBRM is favoured over PRM for modelling the counts. In addition, the authors discovered that both PRM and the Bayesian quantile regression models provide a more accurate estimate of the total counts. More specifically, it was found that the PRM underestimates the actual counts by 0.69% and the Bayesian quantile regression model overestimates the actual counts by 0.79%. On the other hand, the authors noted that NBRM overestimates the actual counts by 3.65%. Based on the results of their study, the authors explain that while the use of quantile

regression models was originally meant for continuous variables, a little transformation to the data by adding a random uniform distribution can enable quantile regression to perform well on the claim count data set.

Fuzi *et al.* (2016) explain that the effect of some covariates was less (or more pronounced) in the upper tail (or lower tail) of their distributions which allowed the authors to have a peek at the shape of the covariates as opposed to the mean regression models from which one can only make conclusions based on central tendency. As such, the authors advise that quantile regression model can be a great extension to the typical mean regression models, thus giving alternatives that can be used to understand claim count data. Fuzi *et al.* (*ibid*) caution that the application of Bayesian quantile regression model in their study is limited to insurance count data which does not involve zero-inflation. As such, the authors advise that careful consideration and the use of other modified approaches may be required for applying the Bayesian quantile regression model to the zero-inflated count data. The Bayesian quantile regression model is not considered in this present study because modification of the quantile model is beyond the scope of this study which only considers mean regression models.

It is worth noting that, for all studies reviewed hereto, only the study by Fuzi *et al.* (2016) has considered the Bayesian quantile regression model as a potential alternative to the commonly used count data models. As such, further research may have to consider quantile models. The effect of sample size on the efficiency of the proposed models was not the interest of the study by Fuzi *et al.* (*ibid*), which differentiates it from this present study. Another difference between the study under review and this present study is that more comparison criteria are applied when selecting models as a way of minimising selection bias. It is worth noting that, other count data models such as zero-inflated and hurdle models were not considered in the study by Fuzi *et al.* (*ibid*), hence this present study extends the scope of the study under review.

Zhan *et al.* (2015) conducted a study to investigate the efficiency of Multivariate Poisson lognormal (MVPLN) model by comparing it to the univariate Poisson-lognormal (PLN) model, PRM and NBRM. The authors used two data sets namely: census level pedestrian–vehicle crash data from New York for the years 2002 to 2006 with 2183 cases and the highway crash data from Washington for the period 1990 to1994 with 275 cases. Zhan *et al.* (*ibid*) explain that since the data from Washington had fewer observations and variables, it was used as a supplement in their study.

Using the data from New York, Zhan *et al.* (2015) modelled the count of pedestrian-vehicle crashes with 16 exploratory variables such as signalized intersections, length of roads in tract and subway stations in tract, to mention a few. The authors also used the data from Washington to model the count of highway injuries using nine exploratory variables which included inter-alia the highway segment length, the log of average annual daily travel per lane and maximum grade difference in the segment. MVPLN, PLN, PRM and NBRM were applied to each of the two data sets. The Deviance information criterion (DIC) and the log-likelihood were then used for comparing the four models. The authors explain that the final models were selected based on the lowest values of DIC and log-likelihood. The results of the study by Zhan *et al.* (*ibid*) revealed that MVPLN outperformed the PLN, PRM and NBRM in both of the two data sets.

The study by Zhan *et al.* (2015) differs from all the studies that have been reviewed hereto because it assessed the performance of MVPLN and PLN using DIC which was not done in other studies. MVPLN and PLN as well as DIC as a comparison criterion are also not considered in this present study because they are beyond its scope. However, this present study extends the scope of the study by Zhan *et al.* (*ibid*) by considering zero-inflated and hurdle models as well as using the Vuong's test, LRT and Mc Fadden's R^2 as model comparison criteria. This present study also deviates from the one by Zhan *et al.* (2015) because it uses ten

data sets with the same variables in order to control for any hallo effect that may arise from comparing models under different data sets with different variables.

The study by Zamani and Ismail (2010) compared the Negative Binomial-Lindley (NB-L) model with PRM and NBRM using two insurance claims data sets which were obtained from Klugman *et al.* (2008). The authors did not give a full description of the independent variables that were used when modelling the count of insurance claims. Zamani and Ismail (*ibid*) used the log-likelihood and Chi-square tests as model selection criteria. The authors found that based on the log likelihood and $p - value$, the NB-L distribution provided the best fit for the data for both data sets. The study by Zamani and Ismail (*ibid*) differs from all other studies reviewed in this present study thus far in that it considered NB-L. NB-L is also not considered in this present study because the present study only compares the popularly used count data models. The present study extends the scope of that of Zamani and Ismail (*ibid*) because it compares the models under more than two data sets and it also considered the zero-inflated and hurdle models. The present study also bears an advantage of using more comparison criteria as a way of minimising the selection bias.

Moghimbeigi *et al.* (2008) conducted a study to confirm the efficiency of multilevel (two fold random effects) ZINB regression model which is designed to incorporate random effects in order to account for data dependency and over-dispersion. The authors used the Decay-missing-filled tooth (DMFT) index data which comprise the sum of the decayed, missing, and filled surfaces of the permanent and primary teeth of children aged 12 years. The other models that were considered in the study by Moghimbeigi *et al.* (*ibid*) are PRM, NBRM, ZIP and ZINB. However, the authors noticed that the data were over-dispersed (significant dispersion parameter from NBRM) and zero-inflated (score test). As such, the authors concluded that ZINB is more appropriate to model the data than the PRM and NBRM. Moghimbeigi *et al.*

(*ibid*) then compared ZINB, multilevel ZIP and multilevel ZINB using the log-likelihood and the scale parameter.

Moghimbeigi *et al.* (2008) found that the multilevel ZINB provides a better fit than multilevel ZIP model because the former had a lower value of the log-likelihood. In addition, the scale parameter estimate was significant, which led Moghimbeigi *et al.* (*ibid*) to conclude that there was substantial over-dispersion in the non-zero counts. Based on the results of the study, the authors concluded that multilevel ZINB is more efficient than other proposed models in their study because the method can provide insight into the source of excess zeros and the apparent heterogeneity and relative dependency inherent in the data structure. The study by Moghimbeigi *et al.* (*ibid*) differs from all other studies reviewed hereto because it considered multilevel ZIP and multilevel ZINB which were never considered in the previously reviewed studies. Due to a limited scope, this present study does not consider multilevel models but all other models that were considered in the study by Moghimbeigi *et al.* (*ibid*) are also applied in this present study. Sample size variations were not the interest of the study by Moghimbeigi *et al.* (*ibid*) which differentiates it from this present study.

Loeys *et al.* (2012) compared six count data models namely: PRM, NBRM, ZIP, ZINB, PHM and NBHM. The data analysed in the study by Loeys *et al.* (*ibid*) is from a sample of 387 subjects who participated in the assessment of the extent of unwanted pursuit behaviour (UPB) perpetrations displayed since the time a couple broke up. The authors used the six proposed models to fit the count of UPB occurrences using a binary indicator for education level and a continuous measurement for the level of anxious attachment in the former partner relationship as predictor variables. Loeys *et al.* (*ibid*) used the LRT and Vuong's tests in comparing the nested and non-nested models respectively. AIC was used in comparing all of the six models.

Loeys *et al.* (2012) found that AIC of the zero-inflated and hurdle models were very close where the lowest AIC was observed with the ZINB and NBHM. In addition, Loeys *et al.* (*ibid*) used the predicted frequencies from each of the proposed models where the estimated frequencies from the hurdle models were found to be almost identical to zero-inflated models. The predicted frequencies also informed the authors that the simple PRM cannot account for the large percentage of zero observations; the ZIP can address this lack of fit for the zero observations but it was noticed that it fails to capture the non-zero frequencies properly. On the other hand, the authors noticed that, based on the predicted frequencies, the ZINB distribution fits the data best.

Loeys *et al.* (2012) concur with Rose *et al.* (2006) that the choice between the hurdle and zero-inflated models should be based on the aim and endpoints of the study. The authors remarked that if the goal of fitting the model is prediction, it is not important which modelling framework is used because predictions are almost identical. On the other hand, Rose *et al.* (2006) and Loeys *et al.* (*ibid*) advise that if the goal is inference, model choice should be related to the study goal. The study by Loeys *et al.* (*ibid*) is similar to the present study in that they both consider the six commonly used count data models which are PRM, NBRM, ZIP, ZINB, PHM and NBHM. Both the study by Loeys *et al.* (*ibid*) and the present study use the LRT and Vuong's test in comparing the nested and non-nested pairs respectively and they both use AIC as a model selection criterion. On the other hand, this present study extends the scope of the study under review by considering other selection criteria such as BIC and Mc Fadden's R^2. The study by Loeys *et al.* (2012) is similar to most studies in literature in that it used one data set, whereas this present study aims to explore ten data sets.

Guikema and Goffelt (2008) conducted a study to demonstrate the usefulness of the Conway-Maxwell Poisson (COM-P) model for modelling data that are either under-/ over-dispersed.

The authors compared COM-P, PRM and NBRM by fitting the count of the power cuts on each of the 648 circuits in a power distribution system. The authors clarify that the data were modified to simulate the under-dispersion scenario. The six predictor variables used in the study by Guikema and Goffelt (*ibid*) include the time since the first recorded time at which the trees along each circuit were trimmed and the miles of overhead line that each circuit contains, to mention a few.

The AIC and log-likelihood were used in comparing the three models that were considered in the study by Guikema and Goffelt (2008) for both the under-dispersed and over-dispersed data sets. The findings of the study by Guikema and Goffelt (*ibid*) revealed that when the data are over-dispersed, NBRM and COM-P fit the data better than the PRM. The authors noticed that the log-likelihood of NBRM and COM-P were nearly identical and concluded that the latter does not provide a significant benefit over the goodness of fit for the standard NBRM. The authors advise that the NBRM reduces to PRM when the data are under-dispersed; therefore, the comparison of models for the data that are under-dispersed was limited to COM-P and PRM only. The findings of the study by Guikema and Goffelt (*ibid*) showed that COM-P performs better than PRM for the under-dispersed data in terms of lower values of the log-likelihood and AIC.

Guikema and Goffelt (2008) introduced yet another count data model which was never considered in the studies that have been reviewed hereto. The study by Guikema and Goffelt (*ibid*) was confirmatory in that it intended to confirm the usefulness of COM-P in modelling count data with the hypothesis that the said proposed model will out-perform NBRM and PRM. On the hand, this present study is exploratory in that it intends to explore the impact of sample size on the efficiency of some well-known count data models without any hypothesis in place. This present study extends the scope of the one by Guikema and Goffelt (*ibid*) by also

considering other model comparison criteria as well as many other count data models. This present study is limited to over-dispersed data because it does not intend to exclude NBRM, ZINB and NBHM from the comparisons which would be the case if under-dispersion is considered as well.

Xiao *et al.* (2015) conducted a study to model the count of fire occurrences using the following predictor variables: monthly maximum temperature (Tmax), monthly mean temperature (T), monthly mean relative humidity (H), monthly mean wind speed (S), monthly maximum wind speed (Smax), monthly precipitation (P) and monthly evaporation (E). The authors applied PRM, NBRM, ZIP, ZINB, PHM and NBHM. However, each model was used in two forms namely: fixed and mixed forms. For the fixed models, PRM, NBRM, ZIP, ZINB, PHM and NBHM were used in their original form as described in Chapter 3 of this current study. In order to form mixed models, Xiao *et al.* (*ibid*) added one random-effect parameter to the intercept of PRM and NBRM and two random-effect parameters to zero-inflated and hurdle models. As such, the study by Xiao *et al.* (*ibid*) considered six count data models.

The models in the study by Xiao *et al.* (2015) were compared using AIC, BIC and the plot of differences between predicted and observed probabilities against count class. The authors found that the AIC and BIC of negative Binomial mixed models (NBRM, ZINB, NBHM) were considerably smaller than those of Poisson mixture models (PRM, ZIP, PHM). It was also found that for both the fixed and mixed models, PRM overestimated the zero counts whereas the other ten models almost accurately estimated the zero counts. The authors also gathered that the mixed models generally minimised the error between the actual and estimated probabilities better than the fixed models with the negative Binomial mixed models having smaller error than the Poisson mixed models. Xiao *et al.* (*ibid*) concluded that the negative Binomial mixed-effects model performed better than the other eleven models (including ZINB

mixed model and NBHM mixed model) in modelling the zero-inflated and over-dispersed occurrences of wild fires.

The study by Xiao *et al.* (2015) is similar to the current study and many other studies reviewed hereto in that it explored the performance of the six commonly used count data models which are: PRM, NBRM, ZIP, ZINB, PHM and NBHM. On the other hand, the mixed-effects models were never considered in the studies that have been reviewed hereto. The study by Xiao *et al.* (*ibid*) was able to create mixed-effect models by considering the random effects among counties. On the other, the information about the counties is not provided for the data used in this current study, hence only the fixed models are considered in it. However, the current study extends the scope of the study by Xiao *et al.* (*ibid*) by considering different sample sizes and other model comparison criteria.

Spriensma *et al*. (2013) embarked on a study to compare three count data models namely: a linear mixed model, Poisson mixed model and a two-part mixed model which the authors identified as a Binomial/Poisson mixed distribution model. All the three models were mixed because their equations included random intercepts and random slopes. The authors explain that the general idea behind the Binomial/Poisson mixture is that the target variable has a Binomial distribution for the zero versus non-zero part and a Poisson distribution for the non-zero part. The dependent variable of interest was hypoglycaemic events while education level was used as the predictor variable. In order to compare the models, Spriensma *et al.* (*ibid*) used BIC and the mean square residual (MSR) as well as the accuracy of the models which was assessed using the scatter plots of observed versus predicted values.

The results of the study by Spriensma *et al*. (2013) revealed that the two-part joint mixed model (Binomial/Poisson) gave the lowest values of BIC and MSR, hence the authors concluded that the Binomial/ Poisson model has the best fit for the data under their study. In addition, the

authors noticed that the Binomial/Poisson model clearly performed best in terms of accuracy when assessing the scatter plot of the actual versus predicted values, especially in correctly predicting the zero-counts. The second best model from the three was found to be the Poisson mixed model.

The studies by Spriensma *et al.* (*ibid*) and Xiao *et al.* (2015) are similar in that they both considered mixed-models, but the former differs from the latter and all other studies reviewed hereto in that it considered the two-part mixed model which assumed that the target variable has a Binomial/ Poisson distribution. The current study is similar to that of Spriensma *et al.* (2013) because the former used BIC and MSE, which is similar to MSR, used in the latter. On other hand, the current study extends the scope of the study by Spriensma *et al* (*ibid*) and many other studies reviewed hereto by considering more model comparison criteria and assessing the efficiency of count data models under different sample sizes. The current study only focuses on the most commonly used count data models with the aim of assessing their efficiency without re-parameterization which has failed to identify the ideal models in previous studies. As such, the mixed models are not considered in the current study.

In their study, Almasi *et al.* (2016) modelled the decay-missing-filled tooth (DMFT) index on 9-year-old children using variables such as region (rural or urban), gender and the mother's education (in years), to mention a few. The authors used the Van der Broek's test and the likelihood Chi-square tests in confirming zero-inflation and over-dispersion in the outcome variable. Almasi *et al.* (*ibid*) compared the performance of ZIP, ZINB, ZIGP and their multilevel forms (MZIP, MZINB and MZIGP). The authors explain that multilevel or random effect models are crucial tools for research designs where participant data are organised at more than one level or in other terms nested data. As such, Almasi *et al.* (*ibid*) considered multilevel models for their study because the DMFT index involves teeth nested within different mouths.

The models in the study by Almasi *et al.* (2016) were compared using the goodness of fit test (GOF) based on the log-likelihood Chi-square, the score test and the Vuong's test. The results of the study by Almasi *et al.* (*ibid*) showed that even when using various dispersion parameters, through Monte Carlo simulations, MZIGP is more efficient in modelling the DMFT index even when the count data are under-dispersed. As discussed in the previous paragraphs of this chapter, the current study does not consider any other model other than the six commonly used count data models which are PRM, NBRM, ZIP, ZINB, PHM and NBHM. As such, the multilevel models used in the study by Almasi *et al.* (*ibid*) are not considered in the current study.

The exclusion of the multilevel models in the current study is also based on the lack of information regarding the nesting of the DurationOfMarriage. In addition to the methods used in comparing the models in the study by Almasi *et al.* (*ibid*), the current study also considers the Mc Fadden's R^2, AIC and BIC to mention a few. The role of sample size on the efficiency of the count data models proposed in the current study remains to be the key aspect of interest which was not considered in the study by Almasi *et al.* (*ibid*).

2.4 Trends and Gaps Identified in Literature

Literature shows that several studies compared numerous count data models using different data sets. 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 models. However, there is no count data model that was found to be generally ideal. Several authors, including Rose *et al.* (2006) and Hoef and Boveng (2007), caution that the choice of the model is dependent on the theoretical and/or scientific knowledge of the data being modelled. Most previous studies such as those by Yip and Yau (2005), Famoye and Singh (2006), Rose *et al.* (2006), Burger *et al.* (2009) and Mei-Chen *et al.* (2011) compared PRM, NBRM, ZIP, ZINB,

PHM and NBHM to other count data models. PRM, NBRM, ZIP, ZINB and PHM can therefore be referred to as the most commonly used count data models in literature. The models compared to the most commonly used models in previous studies which are not considered in the current study due to a limited scope include; QPRM, FMNB-2, ZIGB, MZIGP and ZIDP which were considered by Yip and Yau (2005), Hoef and Boveng (2007) and Park *et al.* (2010), to mention a few.

2.5 Theoretical Framework

The current study focuses on the most popular models which were considered by most authors in previous studies namely: PRM, NBRM, ZIP, ZINB, PHM and NBRM. In addition, the general intent of the current study is to better the efficiency of the proposed count data models without repeatedly developing new ones. This is because the iterative development of new models has not been able to assist researchers in identifying the ideal count data. The most common criteria for comparing count data models in literature are AIC, Vuong test, goodness of fit tests and generalised Pearson Chi-square as, discussed by authors such as Yip and Yau (2005), Rose *et al.* (2006), Hoef and Boveng (2007), Burger *et al.* (2009) and Mei-Chen *et al.* (2011). These criteria are adopted in the current study and are discussed in detail in the Chapter 3. The purpose of implementing numerous comparison criteria is to avoid selection bias which may arise when using only one comparison criteria.

Of the studies reviewed, only Park *et al.* (2010) examined the efficiency of count data models under different sample sizes. However, unlike the current study, Park *et al.* (2010) did not aim at determining the minimum sample size required for count data models to be more efficient when data exhibits over-dispersion and excess zeros, but the main focus area of the study by Park *et al.* (*ibid*) was the bias associated with the posterior mean and median of dispersion parameters in FMNB-2. The general expectation in the current study is that the six count data

models under study will become more efficient as the sample size increases. This is because no study has considered sample size as a way of improving the PRM when the assumption of equi-dispersion is violated. The effect of sample size has not been of interest to previous studies despite a well-known fact that assumptions such as linearity, normality and homogeneity may not be an issue when sample sizes are large (Russell, Lisa and Russell, 2008; Bower, 2013 and Little, 2013).

None of the studies reviewed hereto considered marriage data, hence leaving a gap in literature which this current study intended to fill. However, it is worth noting that the main focus of the current study is on the models and not on DurationOfMarriage and the variables used as predictors of such. Although the current study is interested in the models, it does not preclude that the models should specified in line with the theory of modelling DurationOfMarriage. As such, several literatures have been interrogated (Holman, 2006, Reis and Sprecher, 2009 and Cox and Demmitt, 2013) to ensure that the variables included as predictors of marriage are in line with the theory of marriages and divorces. It is worth noting that since there are no previous studies that were focusing on sample size as a potential way of improving the efficiency of count data models, there is no framework guiding the comparison of such models under different sample sizes. As such, this study formulates a comparison framework comprising two phases: the within-sample and between-sample comparison which are described in more detail in Chapter 3.

Chapter 3

Methodology

3.1 Introduction

This chapter discussed the data and statistical methods implemented in this study. As discussed in the previous chapters, this study compared the efficiency of count data models under different sample sizes. The general objective was to determine whether sample size has an impact on the efficiency of widely used count data models namely: Poisson Regression Model (PRM), Negative Binomial Regression (NBRM), Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB), Poisson Hurdle Model (PHM) and Negative Binomial Hurdle Model (NBHM). The ideal model per sample was selected based on its ability to address zero-inflation and under- or over- dispersion and its goodness of fit. The chapter provides a review of methods and materials that were applied to achieve the objectives that were discussed in Chapter 1 of this study.

The remaining sections of Chapter 3 are organised as follows: Section 3.2 describes the data analysed in this study. Section 3.3 discusses the approaches used in ensuring that the assumptions of count data models are valid in this study. Section 3.4 gives an overview of generalised linear models and Section 3.5 explains the mathematical parameter estimation processes for count data models. Section 3.6 explains how the proposed models are compared and how the ideal model is selected in this study. A summary of this chapter is in Section 3.7.

3.2 Description of the Data

3.2.1 Actual data

The data used in Part A of this study was sourced from Data First and it is available at https://www.datafirst.uct.ac.za. The chosen data sets were for the periods of 2010 ($N = 22936$) and 2011 ($N = 20980$). These data were chosen on availability and also shared the methodology of collection as opposed to other data sets collected prior to 2010. The data were collected by Statistics South Africa. The categorical variables used in this study are: MaleRace, FemaleRace, MaleOccupation, FemaleOccupation, MaleStatus (Marital status of husband), FemaleStatus (Marital status of wife), MaleNoTimesMarried (Male number of times married), FemaleNoTimesMarried (Female number of times married), Solemnisation, and MarriageType. The continuous variables are MaleAge, FemaleAge, NoOfChildren (Number of children) and DurationOfMarriage (dependent variable). The choice of these variables was embedded on literature which has identified these selected socio-demographic variables of the couples as significant predictors of marriage life.

The data sets used in this study represent all finalised divorce cases for the years 2010 and 2011. More precisely, the Marriages and Divorces Metadata 2010 and Marriages and Divorces Metadata 2011 compiled by Statistics South Africa (2014) explain that the data sets include data collected through two methodologies namely: methodology for civil marriages, customary marriages and civil unions and methodology for Divorce. Statistics South Africa (*ibid*) explains that the electronic data on marriages and civil unions was accessed through the State Information Technology Agency (SITA). The authors explain that the divorce data are based on successful applications for divorce that have been issued with decrees. Statistics South Africa (*ibid*) adds that the data are collected from the courts that deal with divorce matters using a standard structured form (Divorce Form 07-04) prepared by Statistics South Africa in

collaboration with the Department of Justice. It is further explained that once a decree of divorce is issued, the plaintiff completes the form and the registrar of the courts signs and stamps it and the forms are consolidated and regularly mailed to Statistics South Africa's head office. Data were analysed using the Statistical Analysis Software (SAS®) version 9.3, registered to the SAS Institute Inc. Cary, NC, USA.

The data sets for 2010 and 2011 were merged using the MERGE statement of SAS® as recommended by Tilanus (2008) to form a large data set with $N = 43916$. Since the interest of this study was to explore the performance of count data models at different sample sizes, ten random samples were drawn from the merged data set for 2010 and 2011 ($N = 43916$) in multiples of 10% until $100\%\ (N)$. The random sampling was performed using the SQL procedure of SAS® as recommended by Matignon (2007). There was no specific theoretical reason for choosing these sample sizes (at 10% increments), but the general intention was to simulate scenarios in which count data models are applied to different sample sizes.

One may have chosen a 2% increment for instance, but that would results in too much output hence making the interpretation of results cumbersome. On the contrary, if one could choose a wider increment, say 50%, the study may fail to capture a wider range of sample size scenarios. Since the merged data set comprised of data from two successive years, there was a possibility of duplicates in the merged data set. As such, the presence of duplicates was diagnosed and any duplicates were removed from the data set using the NODUP option under the SORT procedure of SAS® as recommended by Li (2013). The author explains that PROC SORT compares all variable values of each observation to the previous observation and if a match is found, then the observation is not written in the output data set.

The samples were randomly selected from the merged data set with the intention of preserving the distributional characteristics of the merged data set such as mean, standard deviation and

frequency proportions of different variables. By so doing, the comparison noise between the variables with respect to the differences in distributional characteristics of samples was minimised. By keeping the means and variances similar across the samples, the study ensured that the severity of the over-dispersion does not differ much across the samples. This preservation of merge data characteristics included the frequency and distribution of missing values which intended to minimise the bias in terms of the severity of excess zeros across the samples. The general intent here was to minimise the hallo effects and ensure that the aspect that significantly differentiate between the samples is sample size, which was the main interest of this study.

3.2.2 Simulated data set

Data used in Part B of this study were generated using Monte Carlo Simulations in SAS® (Wicklin, 2013). Previous studies have implemented Monte Carlo simulations through the use of SAS Macros as well as the RAND and RANDGEN functions of SAS/IML to estimate the parameters of PRM and NBRM directly (Kleinman, 2011 and Wicklin, *ibid*). However, there are no studies which have successfully estimated the parameters of ZIP, PHM, ZINB and NBHM directly using SAS Macros as well as the RAND and RANDGEN functions of SAS/IML. As such, instead of estimating the parameters of PRM, NBRM, ZIP, ZINB, PHM and NBHM directly using Monte Carlo simulations, this study simulated five random data sets then subsequently fitted the six models under study for each data set.

The five simulated data sets differed in sample sizes namely: 50 000, 250 000, 500 000, 750 000 and 1000 000. This study did not find guidelines on the minimum sample size that should be considered when performing simulations. This researcher started simulating the data with 50 000 simulations because the maximum number of observations in the actual data was N=43916. As such, the simulated data was treated as an extension of Part A of this study. Also,

there are no guidelines on the magnitude of the increment between the samples and whether the increments should be equal or not. The five sample sizes were therefore chosen by the researcher who believed that they are sufficient to cover large data sets and to display trends on the efficiency of count data models as the sample size increases.

The simulation of the five data sets was based on the example discussed by Wicklin (2013). The RAND function of SAS/IML was used to perform the simulations from "table" distributions. The process began with computing the frequencies for each variable in the observed data set (N=43916) which was used in Part A of the study. The intention was to retain the observed proportions in the simulated data sets as a way of ensuring that the simulated data sets depicted realistic scenarios.

For both categorical and continuous variables, the proportions (frequencies) from the observed data set are used as entries for specifying the probabilities of selecting each element from a set of K categories. To differentiate between the categorical and continuous variables, the study used the format procedure (PROC FORMAT) of SAS to define the categories for the former and the latter are left in numerical format. An example of the Monte Carlo simulation used in this study is available in Appendix 1. The five sample sizes were generated by varying the value of *n* in the do loop; where *n* is the number of observations. Following the simulation of data sets, the six models were applied to the each of the five data sets and different tests and criteria discussed in subsequent sections were used to make comparisons of the models. The other steps are similar to those in Part A.

3.3 Assumptions of Count Data Models

The basic assumption of PRM is that the variance and the mean should be identical (equi-dispersion), but it is almost impractical to have such a distribution. The assumption of equi-dipersion is tested by comparing PRM with NBRM as discussed in Section 3.5. If the

assumption of equi-dispersion is violated, the methods that are designed to handle under- or over- dispersion are used instead. The assumption of no serious multicollinearity applies to all the models under study. This study uses the variance inflation factor (VIF) to test whether there is serious multicollinearity among the independent variables.

The VIF values are obtained by fitting a linear regression model of the DurationOfMarriage on the proposed independent variables using the REG procedure of SAS®. The use of VIF in detecting multicollinearity was adopted from Belussi and Orsi (2015) who recommend that the VIF should be less than 10 if the variables do not exhibit serious multicollinearity. The variables found to be highly collinear were removed from the data set. The VIF was computed using linear regression through (1) as adopted from Yan (2009).

$$VIF = \frac{1}{1-R_j^2}, \tag{1}$$

where R_j^2 is the coefficient of multiple determination.

3.4 Overview of Generalised Linear Models for Count Data

Jackman *et al.* (2007) emphasise that all count data models belong to the Generalised Linear Models (GLM) family. As such, it is important to briefly discuss the GLMs in this study in order to generally understand their form and application. Jackman *et al.* (2007) explain that all GLMs use the same log-linear mean function which is defined in (2) but make different assumptions about the remaining likelihood. The difference in the assumptions about the likelihood can be observed in Log-likelihood functions of models under study in Section 3.5. The GLM is defined by:

$$\log \mu = x^T \beta, \tag{2}$$

where μ, x^T and β denote the mean, the transpose of a vector of regressors and the parameter vector respectively.

Jackman *et al.* (2007) explain that PRM is a classic GLM, while NBRM is an extended GLM with an additional shape parameter and both the hurdle and zero-inflated models are classified as zero-augmented models. The authors explain that zero-augmented models extend the mean function by modifying the likelihood of zero counts.

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 section s a generalised form of the said ML algorithm known as the Newton-Raphson as explained by SAS Institute (2010). The algorithm updates the parameter vector $\boldsymbol{\beta_r}$ at each iteration with (3).

$$\boldsymbol{\beta_{r+1}} = \boldsymbol{\beta_r} - \boldsymbol{H^{-1}s}\,, \tag{3}$$

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 as (4) and (5) respectively:

$$\boldsymbol{s} = [s_j] = \left[\frac{\partial \mathcal{L}}{\partial \beta_j}\right] = 0 \tag{4}$$

$$\boldsymbol{H} = [h_{ij}] = \left[\frac{\partial^2 \mathcal{L}}{\partial \beta_i \partial \beta_j}\right] = 0 \tag{5}$$

Hilbe (2014) elaborates that the ML iteration terminates when the difference between the old and updated values is less than a specific tolerance value, usually 10^{-6} where values of the estimates are at their ML.

3.5 Model/parameter estimation

This section discusses the steps involved in deriving the proposed count data models (PRM, NBRM, ZIP, ZINB, PHM and NBHM). The probability functions, the expected values and the log-likelihood functions of each of the proposed models are defined. All the models were fitted using the NLMIXED procedure of SAS®.

3.5.1 Poison and Binomial models

PRM is the most basic count data model designed for use when the outcome variable is equi-dispersed. For **PRM**, the study adopted the methods explained by Karlaftis *et al*. (2010), unless otherwise specified. The probability of the couple's marriage lasting for y_i years is given by (6):

$$P(y_i|x_i) = \frac{e^{(-\lambda_i)}\lambda_i^{y_i}}{y_i!}, \tag{6}$$

where x_i denote the i^{th} set of predictors per couple. The parameter λ_i is defined as the expected number of years in marriage which is specified as a function of predictor variables in (7):

$$E(y_i) = \lambda_i = e^{\boldsymbol{\beta X}_i}. \tag{7}$$

In (7), $\boldsymbol{X}_i$ is a vector of predictor variables such as male race, female age and solemnisation, to mention a few and $\boldsymbol{\beta}$ denotes a vector of parameter estimates.

It follows from (7) that:

$$\ln \lambda_i = \boldsymbol{\beta X}_i \tag{8}$$

The PRM is estimated by maximising the Poisson log-likelihood ($\mathcal{L}$) function expressed in (9) (Hilbe, 2014):

$$\mathcal{L}(\beta: y_i) = \sum_{i=1}^{n} \left\{ y_i \ln(x_i'\beta) - e^{(x_i'\beta)} - \ln y_i \right\} \quad (9)$$

Hilbe (2014) explains that the first derivative of $\mathcal{L}$ equated to zero yields the score or gradient of the Poisson log-likelihood for β_i and the second derivative of $\mathcal{L}$ yields the Hessian matrix (H). The author further explains that the ML variance-covariance matrix ($\sum$) is estimated by solving the negative inverse of H equated to zero and that the estimates of standard errors are obtained by taking the square roots of the diagonal elements of $-H^{-1}$. It is worth noting that the ML algorithm was used to derive the parameter estimates for all count data models under study. As such, a more generalised ML algorithm used in count data models is explained in Section 3.4 of this study.

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 (the dispersion parameter). The general probability function for NBRM is defined by (10) and was adopted from Haab *et al.* (2012) reads as:

$$P(y_i|x_i) = \frac{\Gamma\left(y_i+\alpha^{-1}\mu_i^{2-p_i}\right)\alpha^{y_i}\mu_i^{(p_i y_i-2y_i)}\left(1+\alpha\mu_i^{p_i-1}\right)^{-\left(y_i+\alpha^{-1}\mu_i^{2-p_i}\right)}}{\Gamma(y_i+1)\Gamma\left(\alpha^{-1}\mu_i^{2-p_i}\right)}, \quad (10)$$

where Γ is the Gamma function,

$$\mu_i = E(y_i|\boldsymbol{\beta}), \quad (11)$$

which follows a Binomial probability distribution, p_i and α are additional parameters that allow for flexibility in over-dispersion. The remaining parts of the discussion about deriving the NBRM estimates are adopted from Hilbe (2011), unless otherwise specified. The parameter estimates for NBRM were obtained by maximising the $\mathcal{L}$ function in (12) using the ML algorithm discussed under Section 3.4:

$$\mathcal{L}(\mu; y, \alpha) = \sum_{i=1}^{1} y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right). \quad (12)$$

The first order derivatives equated to zero yield the score or gradient estimates of β and α and the second order derivatives equated to zero yield the Hessian matrix for $\alpha,\ \beta$ and $\beta; \alpha$.

3.5.2 Zero-inflated models

ZIP is a special form of PRM which is used when the equi-dispersion assumption holds and zero-inflation is likely to have a significant impact on the results of PRM. Zero-inflated models are meant to be used when data are zero-inflated and the outcome variable is over-dispersed. The equations discussed in this section were adopted from the SAS Institute (2010) unless otherwise specified. The probability density function, mean and variance of DuratioOfMarriage under ZIP are given by (13), (14) and (15) respectively:

$$f(y) = \begin{cases} \omega + (1-\omega)e^{-\lambda}, for\ y = 0 \\ (1-\omega)\ \frac{\lambda^y e^{-\lambda}}{y!}, for\ y = 1,2,3\ \ldots \end{cases}, \quad (13)$$

$$\mu = \mathrm{E}(Y) = (1-\omega)\lambda, \quad (14)$$

$$Var(Y) = (1-\omega)\lambda\ (1+\omega\lambda) = \mu + \frac{\omega}{1-\omega}\mu^2, \quad (15)$$

where ω denotes the zero-inflation probability and λ is the Poisson mean parameter.

The parameter estimates for ZIP were obtained by maximising the $\mathcal{L}$ function in (16) using the ML algorithm discussed in Section 3.4.

$$\mathcal{L} = \begin{cases} w_i \log\left[\omega_i + (1-\omega_i)e^{-\lambda}\right], for\ y_i = 0 \\ w_i \log[(1-\omega_i) + y_i \log(\lambda_i) - \lambda_i - \log(y_i!)], for\ y_i > 0 \end{cases}, \quad (16)$$

where w_i denotes the weight of the observation.

Equations of ZIP and ZINB discussed in this section were adopted from the SAS Institute (2010) unless otherwise specified. The probability density function, mean and variance of DurationOfMarriage under ZINB are given by (17), (18) and (19) respectively.

$$f(y) = \begin{cases} \omega + (1-\omega)(1+k\lambda), for\ y = 0 \\ (1-\omega)\frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)}\ \frac{(k\mu)^y}{(1+k\lambda)^{y+1/k}},\ for\ y = 1,2,3\ \ldots' \end{cases} \quad (17)$$

where k is the negative Binomial dispersion parameter and ω denotes the zero-inflation probability.

$$\mu = \mathrm{E}(Y) = (1-\omega)\lambda, \quad (18)$$

$$Var(Y) = (1-\omega)\lambda\ (1+\omega\lambda+k) = \mu + \left(\frac{\omega}{1-\omega} + \frac{k}{1-\omega}\right)\mu^2. \quad (19)$$

Parameter estimates for ZINB were obtained by maximising the $\mathcal{L}$ function in (20) using the ML algorithm as follows:

$$\mathcal{L} = \begin{cases} \log\left[\omega_i + (1-\omega_i)\left(1+\lambda\frac{k}{\omega_i}\right)\right], for\ y_i = 0 \\ \log(1-\omega_i) + y_i \log\left(\frac{k\lambda}{\omega_i}\right) \\ -\left(y_i + \frac{\omega_i}{k}\right)\log\left(1+\frac{k\lambda}{\omega_i}\right) \\ +\log\left(\frac{\Gamma\left(y_i+\frac{\omega_i}{k}\right)}{\Gamma(y_i+1)\Gamma\left(\frac{\omega_i}{k}\right)}\right), for\ \ y_i > 0 \end{cases} . \quad (20)$$

SAS Institute (2010) elaborate that there are two link functions and linear predictors associated with zero-inflated distributions (ZIP and ZINB) of which one is for the zero inflation probability (ω) and another is for the mean parameter. The parameters ω and λ were modelled with regression models in (21) and (22) respectively:

$$h(\omega_i) = \boldsymbol{z}_i'\boldsymbol{\gamma} \quad (21)$$

$$g(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \quad (22)$$

SAS Institute (2010) explains that the covariates $\boldsymbol{z}_i$ are determined by the model explaining the zero-inflation and the covariates $\boldsymbol{x}_i$ are determined by the model specified. SAS Institute (*ibid*) adds that regression parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are estimated by the ML algorithm discussed in Section 3.4.

3.5.3 Hurdle models

Hurdle models are designed to address both under- or over-dispersion. In order to obtain the zero-truncated forms of the probability density functions (PDF's), this study adopted the methods explained by Stroup (2012). The general form of the zero-truncated Poisson PDF corresponding to **PHM** is given by (23):

$$P(Y) = \frac{e^{(-\lambda_i)}\lambda_i^y}{y_i!(1-e^{-\lambda_i})}. \tag{23}$$

The parameter estimates for **PHM** were obtained by maximising the $\mathcal{L}$ function in (24) using the ML algorithm of the form described in Section 3.4.

$$\mathcal{L} = \begin{cases} \log p_i & for\ y=0 \\ \log(1-p_i) y \log \lambda_i - \lambda_i - \log(y!) - \log\left[1-e^{-\lambda_i}\right] & for\ y = 1,2,\ldots \end{cases} \tag{24}$$

The PDF for the zero-truncated negative Binomial model corresponding to **NBHM** is defined by (25):

$$P(Y) = \frac{\binom{y+\left(\frac{1}{\alpha}\right)-1}{y}\left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^y\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha}}{1-\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha}}. \tag{25}$$

In order to obtain the parameter estimates, this study maximised the $\mathcal{L}$ function of NBHM in (26) using the ML algorithm of the form described in Section 3.4.

$$\mathcal{L} =$$

$$\begin{cases} \log p_i \quad for\ y=0 \\ \log(1-p_i) + y\log\frac{\alpha\lambda_i}{1+\alpha\lambda_i} - \frac{1}{\alpha}\log(1+\alpha\lambda_i) + \log\left\{\binom{y+\left(\frac{1}{\alpha}\right)-1}{y}\right\} for\ y = 1,2,\ldots \end{cases} \quad (26)$$

3.6 Model comparison and selection

This section discusses the methods used in comparing the efficiency of the models in this study which was performed in two stages namely: the within-sample comparison and the between-samples comparison. The within-sample comparison compared models within the same sample size in order to select the best model from that particular sample size. The statistics used at the within-sample comparison stage were the Likelihood Ratio Test (LRT) and Vuong's test (V) for comparing the models relative to over-dispersion or zero-inflation, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and the Mc Fadden's R^2. Rose *et al.* (2006) recommend that the LRT should be used if one model in the pair being compared is nested within the other. On the other hand, the authors recommend that if the models being compared are non-nested, then the Vuong's test should be used.

Butler *et al.* (2011) emphasise that two models are said to be 'nested' if one of the models constitutes a special case of the other model. For example, Freese and Long (2006) elaborate that NBRM reduces to PRM when the dispersion parameter is zero. The criteria used for the within-sample comparison may not be appropriate for comparing models with different sample sizes because the values of such criteria may be dependent of the sample size. For example, Hilbe (2014) explains that BIC increases as the sample size increases and that larger sample sizes affect $-2\mathcal{L}$. This study used the mean square error (MSE), the mean absolute deviation (MAD) and Pearson correlation coefficients of the observed and predicted DurationOfMarriage to compare the models from different sample sizes.

The model with the smallest values of MSE and MAD from the pair under comparison was preferred (Wegner, 2007). In addition, the model with the highest Pearson correlation coefficients of the observed and predicted DurationOfMarriage was preferred (Potts and Elith, 2006). This study performed both the within-sample and between sample comparisons on the actual and simulated data sets. A more detailed explanation of how the comparison criteria were derived is given in Sub-sections 3.6.1 and 3.6.2.

3.6.1 Step1. Within-sample comparison

The within-sample comparison entailed the selection of the best model within a sample size (10%, 20%,…,100%) based on its ability to handle under- or over- dispersion and/or zero-inflation, and started from the most basic or standard model to more complex models. The process began with comparing the basic PRM to ZIP to check whether there was significant zero-inflation in the data, assuming that equi-dispersion exists. When PRM was favoured by most comparison criteria from AIC, BIC, Vuong/LRT and Mc Fadden's R^2, this study assumed that zero-inflation is not likely to be an issue and compared PRM with NBRM stage 2 with ZIP being eliminated. However when ZIP was favoured at stage 1, ZIP and NBRM were then compared in stage 2 to confirm whether it was more significant to model the zero-inflation (modelled by ZIP) than the over-dispersion (modelled by NBRM).

When NBRM outperformed ZIP at stage 2, NBRM and ZINB were compared in stage 3 to confirm whether there was need to model only over-dispersion (modelled by NBRM) or both over-dispersion and zero-inflation (modelled by ZINB). On the other hand, when ZIP outperformed NBRM at stage 2, then ZIP and ZINB were compared at stage 3 to examine whether it was significant to model zero-inflation only assuming that equi-dispersion holds (modelled by ZIP) or it was more important to model zero-inflation and over-dispersion (modelled by ZINB).

When NBRM was selected at stage 3, NBRM and PHM were compared at stage 4. Otherwise when the most favoured model was ZIP, then ZIP and PHM were compared at stage 4. On the other hand, when ZINB outperformed either NBRM or PHM at stage 3, ZINB was compared to PHM at stage 4. The model chosen at stage 4 was then compared to NBHM at stage 5. The models chosen at stage 5 of each sample (10%, 20% up to 100%) were then compared to each other in the between-sample comparison phase which is explained in Section 3.6.2. In order to compare the six models in terms of under- or over- dispersion, this study implemented either the V or the LRT (see Table 3.1).

Table 3.1 Criteria for comparing the proposed models relative to under- or over- dispersion

Models being compared	**Criterion**	**Recommended by**
PRM versus ZIP	Vuong	He *et al.* (2012)
PRM versus NBRM	LRT	Cowen and Kattan (2009)
NBRM versus NBHM	LRT	Cameron and Trivedi (2013)
NBRM versus PHM	Vuong	Encaoua *et al.* (2013)
NBRM versus ZINB	Vuong	Karlaftis *et al.* (2010)
ZINB versus PHM	Vuong	Rose *et al.* (2012)
ZINB versus NBRM	Vuong	Butler *et al.* (2011)

Merkle and Smithson (2013) explain that the role of LRT is to test whether the alternative model (T) or more complex model, for example NBRM is sufficiently better than the null model (V) or the less complex model such as PRM. The general form of the test statistic is given by (27):

$$G^2 = -2\left(\mathcal{L}\left(\widehat{\boldsymbol{\theta}}^V|\boldsymbol{y},\boldsymbol{X}\right) - \mathcal{L}\left(\widehat{\boldsymbol{\theta}}^T|\boldsymbol{y},\boldsymbol{X}\right)\right) \quad (27)$$

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 DurationOfMarriage and its associated predictor variables respectively. A significant p-value of the LRT rejected the null hypothesis indicating that there was significant over-dispersion and a model that was designed to handle such (negative

Binomial-based or hurdle-based) was ideal than the model that assumes equi-dispersion. Rose *et al*. (2006) and Little (2013) suggest that the Vuong (V) statistic should be used in comparing the under- or over- dispersion between the non-nested pairs of count data models. Little (*ibid*) denotes the probability for the predicted outcome given the observed predictor values for model 1 (alternative model, for example, ZIP) by $P_1(Y_i|X_i)$ and the author defines the probability model 2 (null model, for example, PRM) as $P_2(Y_i|X_i)$. Given P_1 and P_2, the Vuong statistic was computed using (28):

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i-\bar{m}_i)^2}}, \tag{28}$$

where m_i for each subject *i* is calculated as:

$$m_i = \log\left(\frac{P_1(Y_i|X_i)}{P_2(Y_i|X_i)}\right). \tag{29}$$

The *V* statistic tests the null hypothesis that the observed outcome value is the same for all models under study. Values of *V* less than -1.96 favour the null model, for example, PRM whereas the values of *V* greater than 1.96 favour the alternative model, for example, ZIP (Greene, 2007). On the other hand, the $|V|$<1.96 yields inconclusive results (Peng, 2014).

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were also used for the within-sample comparison to select the best model based on goodness of fit when the *V* or LRT led to inconclusiveness. The formulas in (30) and (31) were used for computing AIC and BIC respectively and were adopted from Hilbe (2014). The model with the lowest value of AIC and/or BIC was preferred.

$$AIC = -2\mathcal{L} + 2k\ , \tag{30}$$

where $\mathcal{L}$ is the log-likelihood function and *k* is the number of parameters in the model.

$$BIC = -2\log\mathcal{L} + k\log n\,, \tag{31}$$

where n is the sample size.

This study adopted the Mc Fadden ρ^2 also known as the Mc Fadden's R^2 for comparing the proposed models in terms of the variation explained as suggested by Karlaftis *et al.* (2010). The ρ^2 statistic is computed using (32):

$$\rho^2 = R^2 = 1 - \frac{\mathcal{L}(\boldsymbol{\beta})}{\mathcal{L}(\mathbf{0})}\,, \tag{32}$$

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. The model with a bigger value of the Mc Fadden's R^2 between the pair being compared was preferred (Karlaftis *et al.*, 2010).

3.6.2 Step 2. Between-Sample Comparison

The between-sample comparison compared models from different sample sizes using the MSE, MAD and Pearson correlation coefficient of the actual versus the predicted values of DuratioOfMarriage. The MSE of the observed values against the predicted values was also used in comparing the models between samples. The choice of the MSE was adopted from Wegner (2007). The model with the lowest value of MSE was preferred. The MSE was computed using (33):

$$MSE = \frac{\sum(Y_{Ai} - Y_{Pi})^2}{n}\,, \tag{33}$$

where Y_{Ai} is the i^{th} actual value of DurationOfMarriage and Y_{Pi} is the i^{th} predicted value of DurationOfMarriage. The model with a smaller value of MSE was preferred because it minimised the error.

In order to avoid biasness, additional method for determining the error margin of the competing model was used. The MAD was computed using (34) which is adopted from Bajpai (2009):

$$MAD = \frac{\sum_{i=1}^{n}|x-\bar{x}|}{n}, \tag{34}$$

where x is the estimated DurationOfMarriage, $\bar{x}$ is the mean DurationOfMarriage and *n* is the sample size. A smaller value of MAD was preferred.

This study also implemented the Pearson correlation coefficients of the actual versus predicted values of the DurationOfMarriage adopted from Potts and Elith (2006):

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}, \tag{35}$$

where x_i and y_i denote the *i*th observation of the actual and predicted values of the DurationOfMarriage respectively, and $\bar{x}_i$ and $\bar{y}_i$ denote their respective means. Models with large values of the Pearson correlation coefficients of the actual versus predicted values of the DurationOfMarriage were preferred (Potts and Elith, 2006).

In addition to the formal comparison of the efficiency, this study also used a histogram in assessing the frequencies of the observed and estimated values of the DurationOfMarriage for the model selected as most efficient in the between-sample comparison phase. The intention was to visualise the number of correctly predicted frequencies. This study implemented the likelihood-ratio chi-square test in assessing the overall significance of the most preferred model (Stokes *et al.,* 2012). The likelihood-ratio chi-square test statistic was computed using (27) which as previously discussed is adopted from Merkle and Smithson (2013). When used for testing the overall significance of the model, $\widehat{\boldsymbol{\theta}}^V$ and $\widehat{\boldsymbol{\theta}}^T$ in (27) are maximum likelihood estimates of the null or intercept only model and the full model with all parameters included in it respectively. The significance of parameter estimates was tested in this study using the Wald

test which was adopted from Enders (2010) who explains that the Wald statistic (ω) is calculated using (37).

$$\omega = \frac{\hat{\theta} - \theta_0}{\mathrm{var}(\hat{\theta})}, \tag{37}$$

where $\hat{\theta}$ is the maximum likelihood parameter estimate and θ_0 is the hypothesised value. A significant p-value of the Wald test rejects the null hypothesis that the parameter estimate is insignificant.

3.7 Summary

In this chapter, the ten data sets used in this study were discussed in detail and the procedures for deriving the proposed models were presented with formulas explained. Criteria for comparing and selecting the models from within-sample and between-sample were also discussed in detail. The ideal model was generally chosen based on the following criteria: the nature of dispersion of the data (tested using either the Vuong's test or LRT), the lowest values of AIC and BIC, the highest value of Mc Fadden's R^2 well as the Pearson correlation coefficients of the actual versus predicted DurationOfMarriage and the lowest values of MSE and MAD. The significance of parameters was tested using the Wald's test. Chapter 4 presents the results of data analysis.

Chapter 4

Data Analysis and Results

4.1 Introduction

This chapter presents the results of the within and between-sample comparisons of the six count data models under study. The analysis of data was done in line with the methodology discussed in Chapter 3. Part A used actual data and Part B uses Monte Carlo Simulated data. The statistics used in comparing the models in the within-sample comparison phase were the LRT (nested pairs) or Vuong's test (non-nested pairs), AIC, BIC and Mc Fadden's R-square as discussed in the previous chapter. The models selected from the within-sample comparison phase were then compared to each other in the between-sample comparison phase in order to select the most efficient count data model for modelling the DurationOfMarriage.

The statistics used in the between-sample comparison were the Pearson correlation coefficient of the actual versus the predicted values of DurationOfMarriage, MSE and MAD. The remainder of this chapter is structured as follows: Section 4.2 describes the distribution of DurationOfMarriage for all the samples and Section 4.3 discusses the results of the diagnosis of duplicates and multicollinearity tests and the remedial methods. Section 4.4 discusses the within-Sample comparison and selection of models. The between-sample comparison and selection of models is presented in Section 4.5. Section 4.6 discusses the significant predictors of DurationOfMarriage and the summary of this chapter is presented in Section 4.7.

Part A: Comparison of the models using actual data sets

4.2 Distribution of DurationOfMarriage (Part A)

This section explores the distribution of DurationOfMarriage to ensure that the key parameters of dispersion (mean and variance) are kept approximately equal across all the ten samples under study. The practice of keeping the mean and variance approximately constant across all samples was perceived to be beneficial to this study as it was assumed to minimise the noise that may bias the dispersion parameters (α) of the models studied due to differences in the measures of dispersion across samples.

The values of the mean and variance were also used in examining the possibility of over-dispersion in the samples. This section also presents a graphical representation of the distribution of DurationOfMarriage to check for either zero-inflation or deflation. The samples explored in the study were referred to as 10% sample, 20% sample up until 100% sample, where 100% was the biggest sample size which was made up of the merged data set for 2010 and 2011. The n% sample sizes were randomly selected from the 100% sample size while preserving the distributional attributes of the 100% sample size. The sample sizes corresponding to these percentages as well as the percentage of observed zero counts to assess zero-inflation are given in Table 4.1.

Table 4.1 Mean and Variance for DurationOfMarriage

Sample	N	Mean	Variance	% 0's
10%	4392	11.09016	74.5225	4.94
20%	8783	11.16429	74.88954	4.94
30%	13175	11.13935	74.54108	4.93
40%	17566	11.13509	74.78146	5.04
50%	21958	11.10807	74.42185	5.16
60%	25107	11.61158	71.42078	5.17
70%	29311	11.60428	71.36708	5.09
80%	33478	11.59182	71.33857	5.09
90%	37667	11.59062	71.22756	5.14
100%	41881	11.5885	71.15306	5.08

Table 4.1 shows that the mean and variance of DurationOfMarriage were approximately equal across all samples. The variance was about seven times the mean implying that the dependent variable may be over-dispersed. The study therefore implemented either the LRT for nested models or the Vuong's test for non-nested models in order to determine whether or not significant over-dispersion exists in the data. The percentage of zeros in each sample was approximately five, implying that there are not too many observed zeros in the data. The histograms for target count variable DurationOfMarriage were examined in order to give a visual representation of the results in Table 4.1. The histograms were meant to confirm that the distribution of the DurationOfMarriage of marriage was approximately equal across all sample sizes.

Figure 4.1 confirms the results in Table 4.1 that the distribution of DurationOfMarriage was approximately equal across all samples. Therefore, there was no bias in the dispersion parameters that could arise due to differences in the means and variances across the samples. There was no spike at zero; therefore zero inflation is unlikely, but there was a possibility of zero-deflation especially for samples of 80% or above due to a dip at zero in the histograms (Zuur *et al.*, 2009). However, the significance of zero inflation was determined by comparing

zero-inflated models with other count data models in this chapter in the within-sample comparison phase.

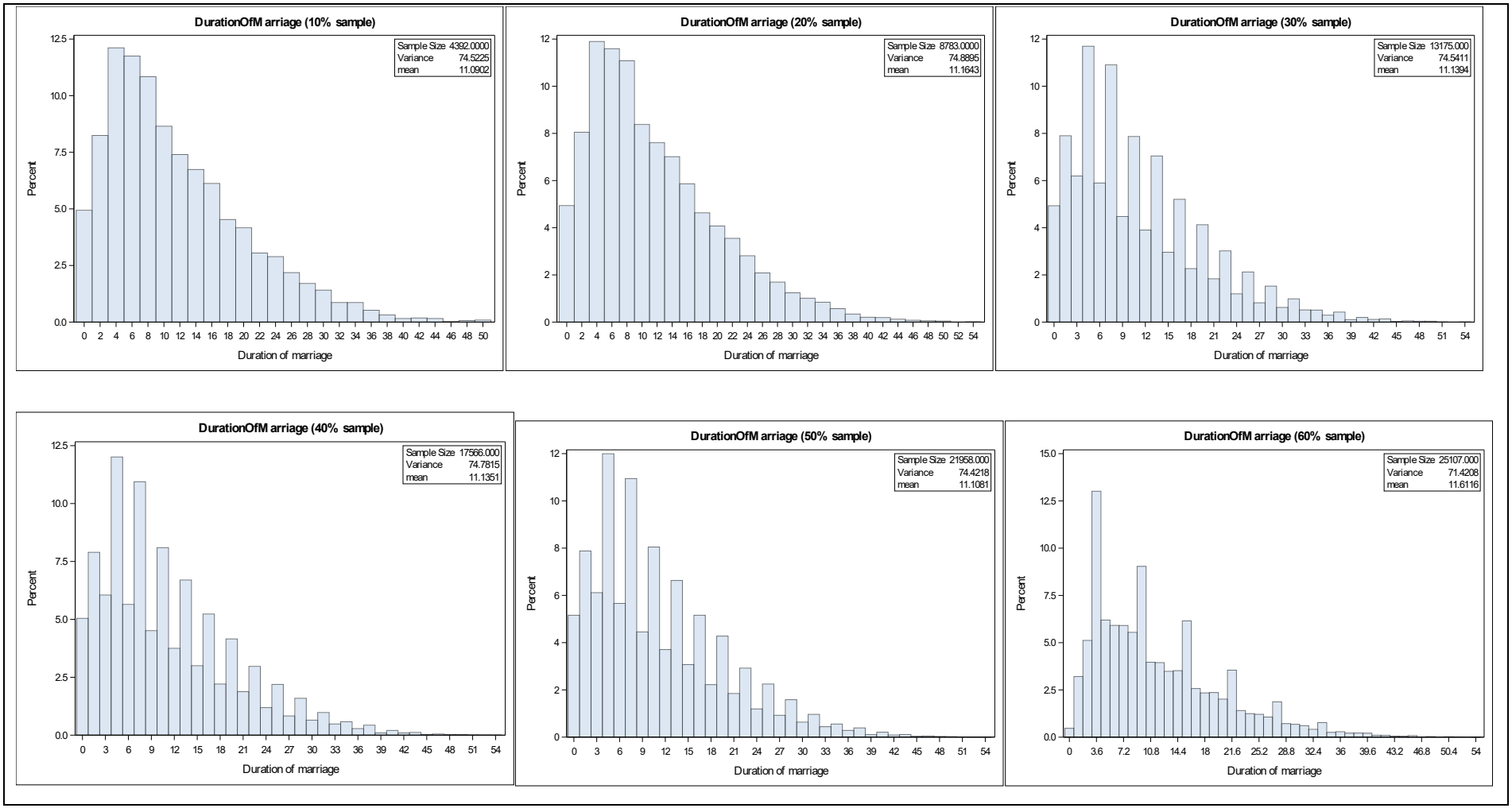

Figure 4.1 Distribution of DurationOfMarriage

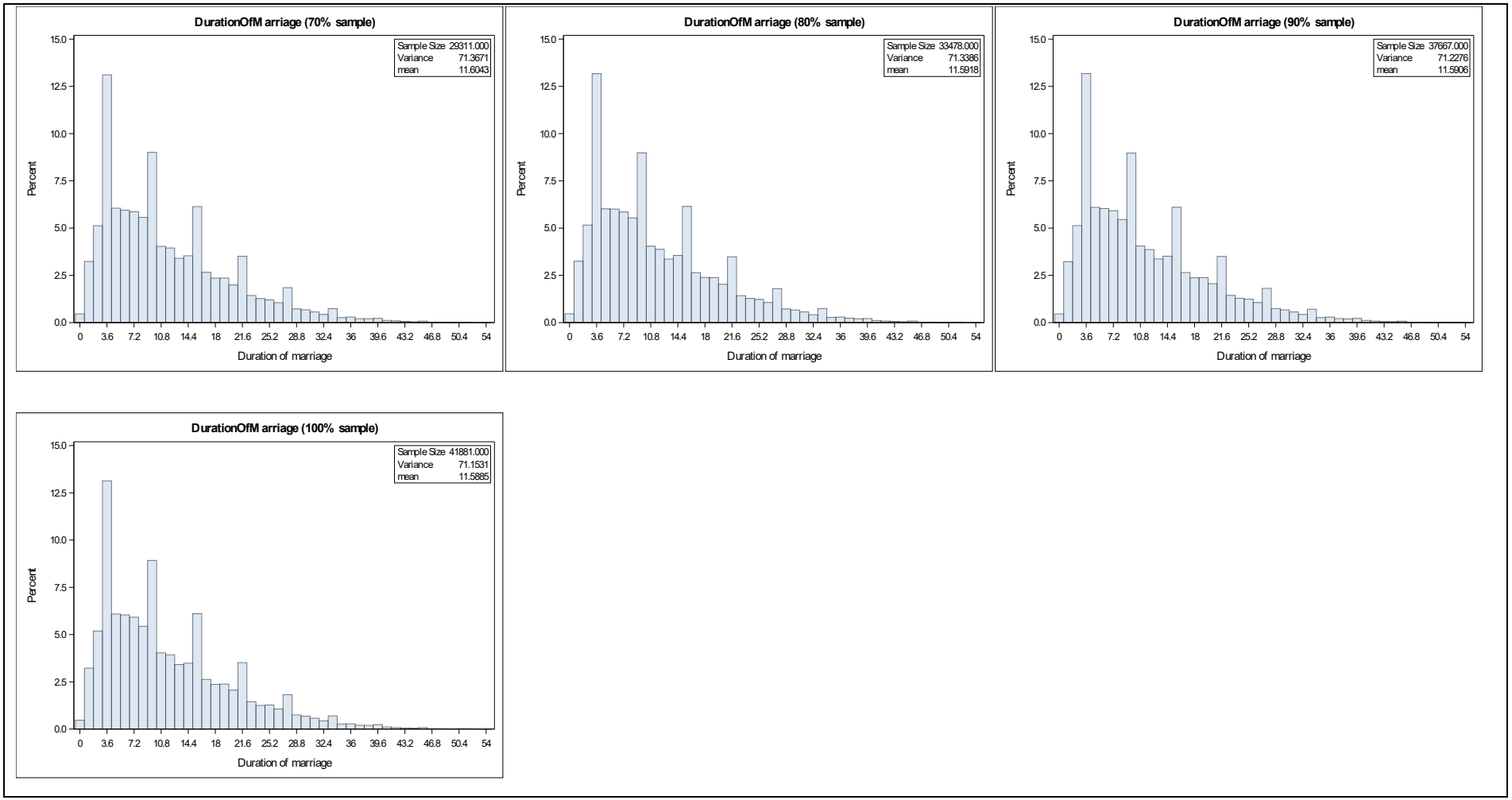

Figure 4.1 Distribution of DurationOfMarriage (continued)

4.3 Diagnosis of duplicates and multicollinearity tests (Part A)

This section examines the data for duplicates because the main data set was formulated from merged data which was collected in consecutive years. Therefore there was a suspicion of repeated subjects in the merged data set. The section presents the results of the variance inflation factor (VIF) to determine whether or not serious multicollinearity exists among the regressors.

This study found it necessary to check for duplicates in data sets because the merged data set (100% sample) comprise of data collected in consecutive time periods hence a suspicion that some subjects may have been repeated. There were no duplicates detected in all the samples under study. Table 4.2 and Table 4.3 present the results of multicollinearity test and the measures implemented in dealing with highly collinear variables.

Table 4.2 Multicollinearity tests using the Variance Inflation Factor (VIF)

Variable	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Male Occupation	1.333	1.329	1.308	1.29	1.288	1.28962	1.29266	1.28665	1.28488	1.28302
Female Occupation	1.353	1.338	1.311	1.292	1.289	1.29094	1.29216	1.28416	1.28430	1.27371
Male Status	4.91	4.508	4.858	5.113	5.037	4.68667	4.47341	4.53578	4.26115	1.42678
Female Status	4.322	4.367	4.864	4.632	4.725	4.46674	4.53435	4.38927	4.37797	1.43710
Male No Times Married	5.178	4.785	5.163	5.352	5.239	4.82521	4.47926	4.57273	4.18724	1.79920
Female No Times Married	4.443	4.499	5.04	4.75	4.856	4.50836	4.51462	4.39469	4.31905	1.80777
MaleAge2	3.202	3.182	3.223	3.223	1.137	3.22262	3.21445	3.19332	3.16173	2.27460
FemaleAge2	3.065	3.116	3.153	3.145	1.133	3.15591	3.14618	3.12137	3.09830	2.24936
Solemnisation	1.321	1.346	1.327	1.332	1.339	1.34440	1.34305	1.34824	1.34604	1.14801
Marriage Type	1.321	1.34	1.313	1.329	1.337	1.33847	1.33382	1.34334	1.34303	1.11899
No of children	1.09	1.091	1.091	1.089	1.075	1.08854	1.08906	1.08687	1.08589	1.05913
Male Race	***16.42***	***18.71***	***19.37***	***17.69***	***18.1***	***18.4979***	***19.5892***	***18.9322***	***18.9153***	***14.5549***
Female Race	***16.47***	***18.8***	***19.45***	***17.74***	***18.13***	***18.5243***	***19.6236***	***18.9693***	***18.9560***	***14.5879***

Male Race and Female Race have $VIF > 10$ for all samples, indicating that the said variables were highly collinear (Belussi and Orsi, 2015). The values of these variables were combined because they were theoretically measuring the same characteristic (race). The VIF values after combining the variables are presented in Table 4.3. $VIF < 10$ Indicates that there was no serious multicollinearity that can significantly hinder the efficiency of the proposed count data models.

Table 4.3 VIF values after combining highly collinear variables

Variable	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
MaleOccupation	1.35	1.343	1.32	1.286	1.297	1.29734	1.29951	1.29330	1.29184	1.28153
FemaleOccupation	1.368	1.348	1.321	1.291	1.296	1.29804	1.29906	1.29092	1.29125	1.27506
MaleStatus	4.886	4.495	4.837	5.186	5.012	4.67649	4.46059	4.52207	4.24812	1.43134
FemaleStatus	4.3	4.363	4.858	4.651	4.707	4.45235	4.51921	4.37412	4.36207	1.44257
MaleNoTimesMarried	5.141	4.757	5.132	5.42	5.217	4.81614	4.46597	4.55731	4.17094	1.81387
FemaleNoTimesMarried	4.414	4.481	5.025	4.781	4.837	4.49045	4.49502	4.37645	4.29908	1.82043
MaleAge2	3.264	3.248	3.28	3.235	1.138	3.28462	3.26738	3.24719	3.21447	2.29959
FemaleAge2	3.124	3.183	3.213	3.171	1.133	3.21773	3.19750	3.17303	3.14835	2.28002
Solemnisation	1.333	1.358	1.337	1.315	1.351	1.35547	1.35278	1.35837	1.35611	1.15094
MarriageType	1.331	1.354	1.324	1.309	1.348	1.34817	1.34248	1.35233	1.35170	1.12120
NoOfChildren	1.087	1.088	1.09	1.081	1.073	1.08693	1.08810	1.08608	1.08536	1.05751
CoupleRace	1.313	1.314	1.298	1.216	1.304	1.31330	1.31495	1.31074	1.31285	1.26240

Table 4.3 shows that for all the samples, $VIF < 10$, which implies that there was no serious multicollinearity after combining the previously highly collinear variables. The findings in this regard were in accordance with the recommendation by Belussi and Orsi (2015). It is worth noting that Female and Male ages were divided by ten (rescaling) and modified to Female Age2 and Male Age2 prior to the analysis. By so doing, the variables Female Age and Male Age were brought to be at par with all other variables whose values were less than ten. This practice of rescaling the parameters was recommended by Kiernan and Tao (2012) as a way of improving optimality of count data models in SAS.

4.4 Within-Sample Model Comparison and Selection of models (Part A)

This section presents the results of an experiment that compared the six count data models of interest using Vuong's test or LRT, AIC, BIC and Mc Fadden's R^2. These are collective criteria used for selecting the model that best fits the data. The model with bigger values of Mc Fadden's R^2 from the ones being compared was preferred (Karlaftis *et al.*, 2010). Moreover, a model with smaller values of AIC and BIC from the ones being compared indicated a better model (Andreff, 2011; and Hilbe, 2014). The Vuong's test and LRT generally tested the hypothesis that the dispersion parameter is zero. The value of the Vuong statistic less than -1.96 favoured the null model while value greater than 1.96 favoured the proposed alternative model. Inversely, the $|V| < 1.96$ yielded inconclusive results (Peng, 2014). The comparison started from basic models that assume equi-dispersion to more complex methods that were designed to handle under- or over- dispersion.

The parameter estimates of all the proposed models are also presented in this section. The significance of the parameter estimates was determined using the p-values corresponding to the Wald' test as recommended by Cameron and Trivedi (2013). For the purpose of this experiment and for uniformity in the number of explanatory variables, all the models were compared. The models include all predictor variables regardless of whether or not they were significant. However, the significant variables were noted as they are important in predicting the DurationOfMarriage.

Table 4.4.1 Parameter Estimates for the Models Implemented in the 10% Sample

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	0.3372(<.0001)*[2]	0.3939(<.0001)*	7.6465(0.9578)	0.003922(0.9681)	0.04951(0.6016)	-1.2230(0.5594)	0.5235 (<.0001)*	-0.2974 (0.7938)	0.06112 (0.5216)	-0.6582(0.6869)
b1	0.000542(0.8069)	0.000380(0.8642)	-0.04787(0.4318)	0.003462(0.3374)	0.002875(0.4055)	-0.04041(0.5867)	-0.00267 (0.0027)*	-0.00384 (0.9332)	0.002595 (0.4551)	-0.04119(0.4536)
b2	-0.00074(0.7207)	-0.00001(0.9960)	0.04945(0.3810)	0.000386(0.9089)	0.001137(0.7255)	0.06182(0.3774)	0.001043 (0.2070)	0.01201 (0.7789)	0.001167 (0.7196)	0.04210(0.4098)
b3	-0.108(<.0001)*	-0.09277(0.0007)*	0.4893(0.2771)	-0.1007(0.0099)*	-0.08701(0.0218)*	0.9099(0.1600)	-0.02765 (0.0132)*	0.2438 (0.4942)	-0.08328 (0.0318)*	0.6485(0.1624)
b4	-0.1833(<.0001)*	-0.1765(<.0001)*	9.7050(0.9465)	-0.1783(<.0001)*	-0.1735(<.0001)*	0.6375(0.5287)	-0.05507 (<.0001)*	0.3031 (0.3771)	-0.1728 (<.0001)*	1.0982(0.3003)
b5	-0.2459(<.0001)*	-0.2478(<.0001)*	0.2201(0.7544)	-0.2653(<.0001)*	-0.2667(<.0001)*	-0.2786(0.8119)	-0.2826 (<.0001)*	0.02319 (0.9656)	-0.2772 (<.0001)*	-0.1228(0.8744)
b6	-0.1127(0.0011)*	-0.1257(0.0004)*	-19.2495(0.9470)	-0.1310(0.0104)*	-0.1419(0.0037)*	-1.4210(0.4314)	-0.3190 (<.0001)*	0.1465 (0.7462)	-0.1485 (0.0027)*	-1.9377(0.3355)
b7	0.1948(<.0001)*	0.1882(<.0001)*	-0.4055(0.2149)	0.2325(<.0001)*	0.2213(<.0001)*	-0.8954(0.0299)	0.1962 (<.0001)*	-0.5962 (0.0264)	0.2257 (<.0001)*	-0.2601(0.3862)
b8	0.4194(<.0001)*	0.4181(<.0001)*	-0.1064(0.7579)	0.4512(<.0001)*	0.4565(<.0001)*	0.6430(0.1565)	0.4033 (<.0001)*	-0.5283 (0.0545)	0.4526 (<.0001)*	-0.3981(0.1965)
b9	0.02288(0.1344)	0.02029(0.1849)	-0.01931(0.9626)	0.02792(0.2641)	0.02843(0.2356)	-0.1260(0.7992)	0.01340 (0.0331)*	-0.3779 (0.2274)	0.02709 (0.2608)	0.04983(0.8948)
b10	-0.02987(<.0001)*	-0.03205(<.0001)*	-0.1341(0.5397)	-0.03076(0.0106)*	-0.03346(0.0037)*	-0.4339(0.1722)	-0.05432 (<.0001)*	-0.1029 (0.4815)	-0.03288 (0.0046)*	-0.06020(0.7467)
b11	0.1473(<.0001)*	0.1415(<.0001)*	-0.5220(0.0214)*	0.1714(<.0001)*	0.1661(<.0001)*	-0.4375(0.1406)	0.1336 (<.0001)*	-1.2395 (<.0001)	0.1655 (<.0001)*	-0.5737(0.0066)*
b12	0.08217(<.0001)*	0.08087(<.0001)*	-0.06060(0.6985)	0.08799(<.0001)*	0.08690(<.0001)*	-0.09821(0.6192)	0.09714 (<.0001)*	-0.09489 (0.4293)	0.08740 (<.0001)*	-0.04257(0.7605)
Alpha				0.1311(<.0001)*	0.1103(<.0001)*				0.1105 (<.0001)*	

[2] *significant at 5%

The values outside the brackets are parameter estimates for the models whereas the values in brackets are probability values (p-values). These values were compared to the 5% level of significance to confirm the statistical significance of each model. It is worth noting that the log-linear part which models the DurationOfMarriage formed part of all the count data models under study. However, the logit part was only applicable to zero-inflated and hurdle models for modelling zero-inflation and deviations from equi-dispersion. Table 4.4.1 shows that, for the log-linear parts, the parameter estimates of b1 (MaleOccupation) and b9 (Solemnisation) were insignificant at 5% level of significance for all models, except PHM, whereas the parameter estimate of b2 (FemaleOccupation) was insignificant in all of the six models. All other variables in the log-linear model were significant in explaining the DurationOfMarriage.

The variable that is significant in explaining the hurdle (logit) part of NBRM, PHM and NBHM is b11 (number of children) but the hurdle part of the ZINB is significantly explained by b7 (maleage2). All dispersion parameters (alpha) for negative Binomial models were greater than 0 and were significant at 5% level indicating that there was significant over-dispersion in the data. The said over-dispersion was confirmed in Table 4.4.2 by comparing the models using the Vuong statistic. For succeeding sections, only a brief interpretation of the tables of parameter estimates was given. The reader may refer to this interpretation of Table 4.4.1 for guidance.

Table 4.4.2 Within-Sample Model Comparison and Selection for 10% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(4.766752)>1.96	ZIP
			AIC	PRM(13444)>ZIP(13178)	ZIP
			BIC	PRM(13518)>ZIP(13325)	ZIP
			Mc Fadden R^2	PRM(0.70869)<ZIP(0.93157)	ZIP
2	ZIP	NBRM	VUONG	V(3.54722)>1.96	NBRM
			AIC	ZIP(13178)>NBRM(12349)	NBRM
			BIC	ZIP(13325)>NBRM (12429)	NBRM
			Mc Fadden R^2	ZIP(0.6937)>NBRM(0.58744)	ZIP
3	NBRM	ZINB	VUONG	V (114.7747)>1.96	ZINB
			AIC	NBRM(12349)>ZINB(12290)	ZINB
			BIC	NBRM(12429)>ZINB(12442)	ZINB
			Mc Fadden R^2	NBRM(0.58744) <ZINB(0.59029)	ZINB
4	ZINB	PHM	VUONG	V(-4.39533)<-1.96	ZINB
			AIC	ZINB(12290)<PHM(13179)	ZINB
			BIC	ZINB(12442)<PHM(13326)	ZINB
			Mc Fadden R^2	ZINB(0.59029) <PHM(0.69367)	PHM
5	ZINB	NBHM	AIC	ZINB(12290)>NBHM(12288)	NBHM
			BIC	ZINB(12442)>NBHM(12441)	NBHM
			Mc Fadden R^2	ZINB(0.59029)>NBHM(0.5895)	ZINB

Table 4.4.2 shows that at Stage 1, PRM was compared to ZIP where all criteria were in favour of ZIP. The Vuong test favoured ZIP over PRM indicating the need to model zero-inflation in the data under the assumption of equi-dispersion. Smaller values of AIC and BIC for ZIP implied that ZIP provided a better fit than PRM. The Mc Fadden's R^2 showed that the ZIP estimated values were closer to the actual DurationOfMarriage than the PRM estimated values. Based on the results of Stage 1, PRM was eliminated from the comparison and ZIP was compared to NBRM at Stage 2. ZIP had a better explanation of the variance in the DurationOfMarriage than NBRM in terms of the Mc Fadden's R^2 but the majority of criteria were in favour of NBRM indicating the need to model either under- or over- dispersion. It is worth noting that alpha was significantly greater than 0 (Table 4.4.1), therefore one may comment that the data are over-dispersed.

NBRM was compared to ZINB in Stage 3 where all statistics favoured ZINB indicating that there may be a need to model over-dispersion and zero-inflation simultaneously. ZINB was compared to PHM in Stage 4 where all but one comparison statistics was favouring ZINB. Stage 5 compared ZINB to NBHM where both selection criteria (AIC and BIC) were in favour of NBHM[3] but ZINB slightly outperformed NBHM in terms of variation explained in the outcome variable (Mc Fadden's R^2)**.** Due to its inconclusive results ($|V| < 1.96$) in comparing NBHM and other models, the Vuong test was **not reported** in all succeeding tables but AIC, BIC and Mc Fadden's R^2 were applied.

The same logic for comparing and selecting the models in the 10% sample was followed for all sample sizes. As such, the interpretation of the steps involved in the within-sample comparison phase for the remaining samples is brief. The reader may refer to this interpretation used in the within-sample comparison of the 10% sample size for a detailed explanation of results in the succeeding within-sample comparisons.

It can be noticed from Table 4.4.2 that for all stages, AIC and BIC agreed in terms of which model was the best. However, the Mc Fadden R^2 contradicted the AIC and BIC in Stage 2, Stage 4 and Stage 5. In addition, the Vuong's test gave inconclusive results at Stage 5 (Not reported). As such, based on the findings, this study considered the AIC and BIC in selecting the best model for the 10% sample size because these selection criteria agreed at each stage. The comparison and selection based on the AIC and BIC values is presented in Table 4.4.3.

[3] The Vuong's test gave inconclusive results ($|V|$<1.96) when comparing NBHM to the other proposed models. The Vuong's test is therefore not used in Stage 5 of each within-sample comparison table in this study.

Table 4.4.3 Model Comparison Based on AIC and BIC only: 10% sample size[4]

MODEL	AIC	MODEL	BIC
NBHM	12288	NBRM	12429
ZINB	12290	NBHM	12441
NBRM	12349	ZINB	12442
ZIP	13178	ZIP	13325
PHM	13179	PHM	13326
PRM	13444	PRM	13518

Table 4.4.3 shows that all Poisson-based models had relatively high AIC and BIC compared to Negative Binomial-based models. For Poisson-based models, ZIP had smaller AIC and BIC than PHM, which in turn had lower AIC and BIC than PRM. On the hand, Table 4.4.3 shows that NBHM had the lowest AIC compared to the other five models under study. In addition, NBRM had the lowest BIC compared to the other five models. As such, both NBRM and NBHM were considered as best models in the within-sample comparison phase for the 10% sample size. Both NBRM and NBHM were considered for the between-sample comparison phase.

[4] AIC and BIC are sorted in ascending order

Table 4.5.1 Parameter Estimates for the Models Implemented in the 20% Sample

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	0.04379(<.0001)*[5]	0.04396(<.0001)*	1.2705(0.0096)*	0.07113(0.9023)	0.06819(0.6698)	1.8742(0.0719)	0.06819(0.6692)	1.8689 (0.0733)	0.06907 (0.7113)	1.1477 (0.1541)
b1	0.001539(0.4308)	0.001538(0.2408)	0.04389(0.1941)	0.002568(0.5460)	0.002440(0.7569)	0.05095(0.2027)	0.002440(0.7565)	0.05090 (0.2021)	0.002459 (0.7979)	0.03984 (0.2070)
b2	0.001438(0.6323)	0.001439(0.9438)	0.04105(0.1299)	0.002390(0.6908)	0.002273(0.4368)	0.04870(0.1646)	0.002273(0.4369)	0.04864 (0.1647)	0.002290 (0.4224)	0.03712 (0.1162)
b3	0.01801(<.0001)*	0.01803(<.0001)*	0.4631(0.5157)	0.02543(<.0001)*	0.02450(<.0001)*	0.8620(0.3957)	0.02450(<.0001)*	0.8639 (0.4022)	0.02510 (<.0001)*	0.4743 (0.2257)
b4	0.01923(<.0001)*	0.01934(<.0001)*	0.5607(0.3621)	0.02678(<.0001)*	0.02574(<.0001)*	0.9341(0.2484)	0.02574(<.0001)*	0.9283 (0.2461)	0.02623 (<.0001)*	0.5423 (0.2468)
b5	0.02751(<.0001)*	0.02754(<.0001)*	0.7906(0.5888)	0.03758(<.0001)*	0.03625(<.0001)*	1.7145(0.3837)	0.03625(<.0001)*	1.7169 (0.3855)	0.03738 (<.0001)*	0.8613 (0.3423)
b6	0.02847(<.0001)*	0.02871(<.0001)*	0.9629(0.5073)	0.03797(0.0002)*	0.03650(<.0001)*	1.7676(0.4057)	0.03650(<.0001)*	1.7552 (0.4056)	0.03740 (<.0001)*	0.9439 (0.3856)
b7	0.008616(<.0001)*	0.008646(<.0001)*	0.2378(0.7903)	0.01551(<.0001)*	0.01473(<.0001)*	0.2807(0.6400)	0.01473(<.0001)*	0.2807 (0.6365)	0.01484 (<.0001)*	0.2205 (0.6994)
b8	0.009031(<.0001)*	0.009078(<.0001)*	0.2499(0.7319)	0.01607(<.0001)*	0.01530(<.0001)*	0.3033(0.3315)	0.01530(<.0001)*	0.3033 (0.3233)	0.01544 (<.0001)*	0.2270 (0.1619)
b9	0.01083(0.0683)	0.01085(0.0462)*	0.3247(0.5462)	0.01808(0.4117)	0.01721(0.3005)	0.3821(0.7655)	0.01721(0.3011)	0.3805 (0.7784)	0.01733 (0.2689)	0.2843 (0.7377)
b10	0.005348(<.0001)*	0.005345(<.0001)*	0.1594(0.9469)	0.008925(0.0001)*	0.008483(<.0001)*	0.2013(0.6098)	0.008483(<.0001)*	0.2019 (0.5716)	0.008551 (<.0001)*	0.1438 (0.5760)
b11	0.004912(<.0001)*	0.004938(<.0001)*	0.1650(0.0061)	0.008225(<.0001)*	0.007841(<.0001)*	0.2052(0.0969)	0.007841(<.0001)*	0.2045 (0.1016)	0.007897 (<.0001)*	0.1521 (0.0004)*
b12	0.003917(<.0001)*	0.003927(<.0001)*	0.1157(0.8830)	0.006601(<.0001)*	0.006279(<.0001)*	0.1403(0.6029)	0.006279(<.0001)*	0.1399 (0.6113)	0.006325 (<.0001)*	0.1015 (0.9626)
				0.1379(<.0001)*	0.1135(<.0001)*				0.3735 (<.0001)*	

[5] *significant at 5%

Table 4.5.1 shows that, for the log-linear parts, b1 (MaleOccupation), b2 (FemaleOccupation) and b9 (Solemnisation) were insignificant at 5% level of significance for all of the proposed models. None of the variables were significant in modelling zero-inflation (logit) for ZINB and the hurdle (logit) part in PHM. The hurdle part of NBHM and the zero-inflation part of ZIP were described by b9 (Solemnisation). All dispersion parameters (α) for the negative Binomial models were significantly different from 0, indicating that the data were overly dispersed. The models were compared in Table 4.5.2.

Table 4.5.2 Within-Sample Model Comparison and Selection for 20% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(7.408155)>1.96	ZIP
			AIC	PRM(27508)>ZIP(26825)	ZIP
			BIC	PRM(27590)>ZIP(26990)	ZIP
			Mc Fadden R^2	PRM(0.70200)>ZIP(0.68774)	PRM
2	ZIP	NBRM	VUONG	V(6.34488)>1.96	NBRM
			AIC	ZIP(26825)>NBRM(25057)	NBRM
			BIC	ZIP(26990)>NBRM (25146)	NBRM
			Mc Fadden R^2	ZIP(0.70200)>NBRM(0.58162)	ZIP
3	NBRM	ZINB	VUONG	V (83.64964)>1.96	ZINB
			AIC	NBRM(25057)>ZINB(24872)	ZINB
			BIC	NBRM(25146)>ZINB(25044)	ZINB
			Mc Fadden R^2	NBRM(0.58162) <ZINB(0.58514)	ZINB
4	ZINB	PHM	VUONG	V(-7.54295)<-1.96	ZINB
			AIC	ZINB(24872)<PHM(26826)	ZINB
			BIC	ZINB(25044)<PHM(26991)	ZINB
			Mc Fadden R^2	ZINB(0.58514) <PHM(0.68773)	PHM
5	ZINB	NBHM	AIC	ZINB(24872)>NBHM(24868)	NBHM
			BIC	ZINB(25044)>NBHM(25039)	NBHM
			Mc Fadden R^2	ZINB(0.58514)<NBHM(0.58426)	ZINB

It can be noticed from Table 4.5.2 that for all stages, the AIC and BIC agreed in terms of which model was the best. However, there were contradictions between other criteria and AIC and BIC in some stages. As such, this study considered AIC and BIC in selecting the best model for the 20% sample size because these selection criteria agreed at each stage.

Table 4.5.3 Model Comparison Based on AIC and BIC only: 20% sample size[6]

MODEL	AIC	MODEL	BIC
NBHM	24868	NBHM	25039
ZINB	24872	ZINB	25044
NBRM	25057	NBRM	25146
ZIP	26825	ZIP	26990
PHM	26826	PHM	26991
PRM	27508	PRM	27590

The results in Table 4.5.3 are similar to the ones in Table 4.4.3. In general, NBHM had the lowest AIC and BIC compared to the other five models under study, hence this model was considered for the between-sample comparison phase.

[6] AIC and BIC are sorted in ascending order

Table 4.6.1 Parameter Estimates for the Models Implemented in the 30% Sample

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	0.4895(<.0001)*[7]	0.5150(<.0001)*	-3.0859(0.0012)*	0.09953(0.0946)	0.1274(0.0251)	-3.9032 (0.0250)*	0.5210 (<.0001)	-1.8628 (0.0344)	0.1296 (0.0238)	-1.8596 (0.0336)
b1	-0.00183(0.1455)	-0.00272(0.0311)	-0.04728(0.1674)	0.000580(0.7846)	-0.00045(0.8208)	-0.07187 (0.0829)	-0.00269 (0.0333)	-0.03968 (0.2055)	-0.00040 (0.8436)	-0.03940 (0.2054)
b2	-0.00089(0.4461)	-0.00012(0.9197)	0.04905(0.1239)	0.000230(0.9069)	0.001192(0.5224)	0.07259 (0.0638)	-0.00012 (0.9205)	0.04563 (0.1167)	0.001055 (0.5736)	0.04519 (0.1178)
b3	-0.04553(0.0032)*	-0.04428(0.0040)*	0.09905(0.7717)	-0.07044(0.0015)*	-0.06727(0.0015)*	0.4849 (0.4680)	-0.04019 (0.0101)*	0.4114 (0.2408)	-0.06236 (0.0041)*	0.3847 (0.2782)
b4	-0.1366(<.0001)*	-0.1247(<.0001)*	0.4820(0.2765)	-0.1543(<.0001)*	-0.1442(<.0001)*	0.9535 (0.3377)	-0.1226 (<.0001)*	0.5479 (0.2194)	-0.1427 (<.0001)*	0.5484 (0.2177)
b5	-0.2589(<.0001)*	-0.2592(<.0001)*	-0.1738(0.7495)	-0.2142(<.0001)*	-0.2216(<.0001)*	-0.9959 (0.4362)	-0.2667 (<.0001)*	-0.5435 (0.3799)	-0.2324 (<.0001)*	-0.5590 (0.3719)
b6	-0.1754(<.0001)*	-0.1857(<.0001)*	-0.6170(0.4112)	-0.1362(<.0001)*	-0.1487(<.0001)*	-1.4072 (0.4596)	-0.1909 (<.0001)*	-0.8130 (0.2954)	-0.1551 (<.0001)*	-0.8121 (0.2947)
b7	0.1874(<.0001)*	0.1893(<.0001)*	0.07161(0.7010)	0.2140(<.0001)*	0.2157(<.0001)*	0.2316 (0.2987)	0.1895 (<.0001)*	0.01414 (0.9354)	0.2169 (<.0001)*	0.03569 (0.8363)
b8	0.4032(<.0001)*	0.4003(<.0001)*	-0.1451(0.4538)	0.4460(<.0001)*	0.4443(<.0001)*	0.03445 (0.8834)	0.4002 (<.0001)*	-0.2868 (0.1087)	0.4451 (<.0001)*	-0.2896 (0.1023)
b9	0.000637(0.9420)	0.000712(0.9354)	-0.06778(0.7800)	0.003757(0.8006)	0.003526(0.8022)	-0.06329 (0.8320)	0.000800 (0.9274)	-0.1101 (0.6159)	0.003208 (0.8207)	-0.1146 (0.5984)
b10	-0.04367(<.0001)*	-0.04367(<.0001)*	0.005520(0.9627)	-0.04369(<.0001)*	-0.04389(<.0001)*	-0.03564 (0.8143)	-0.04372 (<.0001)*	-0.02634 (0.8053)	-0.04399 (<.0001)*	-0.03409 (0.7478)
b11	0.1422(<.0001)*	0.1378(<.0001)*	-0.4065(0.0015)	0.1682(<.0001)*	0.1638(<.0001)*	-0.248 6(0.1195)	0.1377 (<.0001)*	-0.4935 (<.0001)*	0.1637 (<.0001)*	-0.5118 (<.0001)
b12	0.08400(<.0001)*	0.08211(<.0001)*	-0.06491(0.4731)	0.08565(<.0001)*	0.08446(<.0001)*	-0.1158 (0.2980)	0.08236 (<.0001)*	-0.04560 (0.5734)	0.08515 (<.0001)*	-0.03308 (0.6802)
				0.1448(<.0001)*	0.1165(<.0001)*				8.5288 (<.0001)*	

[7] *significant at 5%

Table 4.6.1 shows that, for the log-linear parts, the parameter estimates of b1 (MaleOccupation) were insignificant for all proposed models except for PHM, whereas the parameter estimates of b2 (FemaleOccupation) and b9 (Solemnisation) were insignificant for all the proposed models. The logit parts of ZIP, PHM and NBHM were significantly explained by b11 (number of children) whereas none of the variables were significant in explaining the zero-inflated (logit) part of ZINB. All dispersion parameters (alpha) were significantly greater than 0, indicating that the data were significantly over-dispersed, the dispersion parameter for NBHM was very large as opposed to the 10% and 20% sample sizes.

Table 4.6.2 Within-Sample Model Comparison and Selection for 30% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(9.3474)>1.96	ZIP
			AIC	PRM(41193)>ZIP(39956)	ZIP
			BIC	PRM(41280)>ZIP(40131)	ZIP
			Mc Fadden R^2	PRM(0.701474)<ZIP(0.710633)	ZIP
2	ZIP	NBRM	VUONG	V(7.61848)>1.96	NBRM
			AIC	ZIP(39956)>NBRM(37339)	NBRM
			BIC	ZIP(40131)>NBRM (37434)	NBRM
			Mc Fadden R^2	ZIP((0.710633)>NBRM(0.58382)	ZIP
3	NBRM	ZINB	VUONG	V (100.2176)>1.96	ZINB
			AIC	NBRM(37339)>ZINB(36994)	ZINB
			BIC	NBRM(37434)>ZINB(37176)	ZINB
			Mc Fadden R^2	NBRM(0.58382) <ZINB(0.58796)	ZINB
4	ZINB	PHM	VUONG	V(-9.194521)<-1.96	ZINB
			AIC	ZINB(36994)<PHM(39954)	ZINB
			BIC	ZINB(37176)<PHM(40129)	ZINB
			Mc Fadden R^2	ZINB(0.58514) <PHM(0.688825)	PHM
5	ZINB	NBHM	AIC	ZINB(36994)>NBHM(36990)	NBHM
			BIC	ZINB(37176)>NBHM(37172)	NBHM
			Mc Fadden R^2	ZINB(0.58796)>NBHM(0.58704)	ZINB

It can be noticed from Table 4.6.2 that for all stages, AIC and BIC agreed in terms of which model is the best but other criteria contradicted these statistics at some stages of the comparison. As such, this study considered AIC and BIC in selecting the best model for the 30% sample size because these selection criteria agreed at each stage.

Table 4.6.3 Model Comparison Based on AIC and BIC only: 30% sample size[8]

MODEL	AIC	MODEL	BIC
NBHM	36990	NBHM	37172
ZINB	36994	ZINB	37176
NBRM	37339	NBRM	37434
PHM	39954	PHM	40129
ZIP	39956	ZIP	40131
PRM	41193	PRM	41280

The results in Table 4.6.3 are similar to the ones in Table 4.4.3. In overall, NBHM had the lowest AIC and BIC compared to the other five models under study, hence it was considered for the between-sample comparison phase.

[8] AIC and BIC are sorted in ascending order

Table 4.7.1 Parameter Estimates for the Models Implemented in the 40% Sample

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***logit***
b0	0.5298(<.0001)*[9]	0.5541(<.0001)*	-3.0399(<.0001)*	0.1492(0.0029)*	0.1788(0.0002)*	-3.3609(0.0037)*	0.5616 (<.0001)*	-1.8357 (0.0090)*	0.1798 (0.0002)*	-1.9009 (0.0070)*
b1	-0.00350(0.0007)*	-0.00421(<.0001)*	-0.03016(0.2646)	-0.00136(0.4423)	-0.00204(0.2208)	-0.04744(0.1392)	-0.00419 (<.0001)*	-0.02367 (0.3400)	-0.00204 (0.2229)	-0.02316 (0.3512)
b2	0.001458(0.1317)	0.002070(0.0325)	0.03376(0.1819)	0.002631(0.1102)	0.003238(0.0364)*	0.04238(0.1611)	0.002075 (0.0321)*	0.03003 (0.1955)	0.003211 (0.0393)*	0.03061 (0.1877)
b3	-0.05571(<.0001)*	-0.05116(0.0001)*	0.2418(0.3727)	-0.07952(<.0001)*	-0.07382(<.0001)*	0.4670(0.3231)	-0.04686 (0.0006)*	0.5286 (0.0674)	-0.06796 (0.0004)*	0.5342 (0.0648)
b4	-0.07211(<.0001)*	-0.06207(<.0001)*	0.4209(0.2403)	-0.09431(<.0001)*	-0.08680(<.0001)*	0.6696(0.3237)	-0.05751 (0.0001)*	0.4867 (0.1943)	-0.08136 (<.0001)*	0.5033 (0.1773)
b5	-0.2505(<.0001)*	-0.2514(<.0001)*	-0.2952(0.5092)	-0.2032(<.0001)*	-0.2126(<.0001)*	-1.0488(0.2347)	-0.2592 (<.0001)*	-0.6958 (0.1810)	-0.2244 (<.0001)*	-0.7003 (0.1791)
b6	-0.2561(<.0001)*	-0.2675(<.0001)*	-0.6249(0.3024)	-0.2015(<.0001)*	-0.2134(<.0001)*	-1.1912(0.3424)	-0.2768 (<.0001)*	-0.8027 (0.2222)	-0.2256 (<.0001)*	-0.8135 (0.2143)
b7	0.1829(<.0001)*	0.1899(<.0001)*	0.3342(0.0188)	0.2066(<.0001)*	0.2156(<.0001)*	0.6369(0.0002)*	0.1899 (<.0001)*	0.2650 (0.0453)*	0.2161 (<.0001)*	0.2489 (0.0609)
b8	0.4138(<.0001)*	0.4076(<.0001)*	-0.3033(0.0422)*	0.4559(<.0001)*	0.4488(<.0001)*	-0.2951(0.0938)	0.4076 (<.0001)*	-0.4354 (0.0017)*	0.4508 (<.0001)*	-0.4112 (0.0031)*
b9	-0.00275(0.7055)	-0.00698(0.3390)	-0.3196(0.0919)	-0.00215(0.8635)	-0.00740(0.5301)	-0.4669(0.0387)*	-0.00684 (0.3489)	-0.3201 (0.0637)*	-0.00711 (0.5486)	-0.3201 (0.0643)
b10	-0.04328(<.0001)*	-0.04265(<.0001)*	0.01929(0.8350)	-0.04230(<.0001)*	-0.04223(<.0001)*	-0.02701(0.8106)	-0.04262 (<.0001)*	0.002335 (0.9775)*	-0.04241 (<.0001)*	0.008950 (0.9140)
b11	0.1405(<.0001)*	0.1357(<.0001)*	-0.3879(0.0002)*	0.1665(<.0001)*	0.1614(<.0001)*	-0.2890(0.0230)*	0.1356 (<.0001)*	-0.4801 (<.0001)*	0.1617 (<.0001)*	-0.4761 (<.0001)*
b12	0.07086(<.0001)*	0.06833(<.0001)*	-0.1295(0.0464)*	0.06908(<.0001)*	0.06743(<.0001)*	-0.1498(0.0565)	0.06845 (<.0001)*	-0.1155 (0.0468)*	0.06783 (<.0001)*	-0.1173 (0.0437)*
				0.1523(<.0001)*		0.1197(<.0001)*			8.2833 (<.0001)*	

[9] *significant at 5%

Table 4.7.1 shows that for the log-linear parts, the parameter estimates of b1 (MaleOccupation) were insignificant for both zero-inflated models and for NBHM, whereas the parameter estimate of b2 (FemaleOccupation) was insignificant for all models except for the hurdle models. In addition, the parameter estimates of b9 (Solemnisation) were insignificant for all log-linear parts of the proposed models. For the logit parts, b11 (number of children) was significant in all zero-inflated and hurdle models. As opposed to the 10%, 20% and 30% samples, there were many more significant predictors in the logit models. All dispersion parameters were significantly different from 0, with a very large dispersion parameter for NBHM as in the 30% sample size.

Table 4.7.2 Within-Sample Model Comparison and Selection for 40% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(11.27373)>1.96	ZIP
			AIC	PRM(59710)>ZIP(57632)	ZIP
			BIC	PRM(59803)>ZIP(57817)	ZIP
			Mc Fadden R^2	PRM(0.676186)>ZIP(0.663315)	PRM
2	ZIP	NBRM	VUONG	V(9.37437)>1.96	NBRM
			AIC	ZIP(57632))>NBRM(53831)	NBRM
			BIC	ZIP(57817)>NBRM (53931)	NBRM
			Mc Fadden R^2	ZIP(0.663315)>NBRM(0.55004)	ZIP
3	NBRM	ZINB	VUONG	V (113.2202)>1.96	ZINB
			AIC	NBRM(53831)>ZINB(53241)	ZINB
			BIC	NBRM(53931)>ZINB(53433)	ZINB
			Mc Fadden R^2	NBRM(0.55004) <ZINB(0.555196)	ZINB
4	ZINB	PHM	VUONG	V(-11.21022)<-1.96	ZINB
			AIC	ZINB(53241)<PHM(57630)	ZINB
			BIC	ZINB(53433)<PHM(57814)	ZINB
			Mc Fadden R^2	ZINB(0.555196) <PHM(0.663619)	PHM
5	ZINB	NBHM	AIC	ZINB(53241)>NBHM(51492)	NBHM
			BIC	ZINB(53433)>NBHM(51591)	NBHM
			Mc Fadden R^2	ZINB(0.55520)>NBHM(0,554176)	ZINB

It is evident from Table 4.7.2 that the results of the within-sample comparison are similar to the ones in Table 4.4.3 in which AIC and BIC agrees on the effective models at each stage contradicted, but other criteria contradicts these statistics in some of the stages As such, this study employed AIC and BIC in selecting the best model for the 40% sample size because these selection criteria agreed at each stage.

Table 4.7.3 Model Comparison Based on AIC and BIC only: 40% Sample Size[10]

MODEL	AIC	MODEL	BIC
NBHM	51492	NBHM	51591
ZINB	53241	ZINB	53433
NBRM	53831	NBRM	53931
PHM	57630	PHM	57814
ZIP	57632	ZIP	57817
PRM	59710	PRM	59803

As seen in Table 4.7.3 are similar to the ones in Table 4.4.3 such that all Poisson-based models had relatively high AIC and BIC compared to Negative Binomial-based models. In general, NBHM had the lowest AIC and BIC compared to the other five models under study, hence it was considered for the between-sample comparison phase.

[10] AIC and BIC are sorted in ascending order

Table 4.8.1 Parameter Estimates for the Models Implemented in the 50% Sample

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	1.6267(<.0001)*[11]	1.6676(<.0001)*	-2.5941(<.0001)*	1.7054(<.0001)*	1.7354(<.0001)*	-0.7979 (0.4213)	1.6724 (<.0001)*	-1.7131 (0.0046)*	1.7314 (<.0001)*	-1.7038 (0.0049)*
b1	0.003564(0.0002)*	0.003079(0.0015)*	-0.01693(0.4813)	0.005766(0.0059)*	0.005464(0.0074)*	-0.02799 (0.4634)	0.00309 6(0.0014)*	-0.01514 (0.5125)	0.005516 (0.0073)*	-0.01602 (0.4913)
b2	0.004863(<.0001)*	0.005381(<.0001)*	0.02208(0.3235)	0.004794(0.0138)*	0.005337(0.0049)*	0.02274 (0.5236)	0.005377 (<.0001)*	0.01859 (0.3875)	0.005360 (0.0050)*	0.01825 (0.3994)
b3	-0.02179(0.0764)	-0.01642(0.1797)	0.2488(0.2933)	-0.03756(0.0806)	-0.03302(0.1160)	0.2272 (0.5540)	-0.01277 (0.3042)	0.4443 (0.1134)	-0.02602 (0.2292)	0.4458 (0.1128)
b4	-0.09400(<.0001)*	-0.08713(<.0001)*	0.2997(0.2874)	-0.1059(<.0001)*	-0.1050(<.0001)*	-0.08929 (0.8829)	-0.08368 (<.0001)*	0.4881 (0.1228)	-0.1004 (<.0001)*	0.5183 (0.0967)
b5	-0.2570(<.0001)*	-0.2578(<.0001)*	-0.2649(0.4977)	-0.1985(<.0001)*	-0.2032(<.0001)*	-0.2168 (0.7339)	-0.2650 (<.0001)*	-0.6669 (0.1875)	-0.2209 (<.0001)*	-0.6752 (0.1830)
b6	-0.1794(<.0001)*	-0.1872(<.0001)*	-0.4186(0.3612)	-0.1093(0.0009)*	-0.1145(0.0004)*	-0.4019 (0.6730)	-0.1944 (<.0001)*	-0.7032 (0.1988)	-0.1248 (0.0002)*	-0.6922 (0.1995)
b7	0.1434(<.0001)*	0.1449(<.0001)*	0.05768(0.2983)	0.1300(<.0001)*	0.1323(<.0001)*	0.1298 (0.1872)	0.1450 (<.0001)*	0.02400 (0.6341)	0.1341 (<.0001)*	0.02655 (0.6015)
b8	0.2029(<.0001)*	0.2011(<.0001)*	-0.04819(0.3729)	0.1658(<.0001)*	0.1673(<.0001)*	0.03394 (0.7195)	0.2016 (<.0001)*	-0.07994 (0.1083)	0.1695 (<.0001)*	-0.09009 (0.0700)*
b9	-0.02039(0.0029)*	-0.02881(<.0001)*	-0.4220(0.0109)	-0.02042(0.1716)	-0.03053(0.0359)*	-1.2723 (<.0001)*	-0.02878 (<.0001)*	-0.4989 (0.0016)*	-0.02923 (0.0461)	-0.4983 (0.0017)*
b10	-0.05985(<.0001)*	-0.05994(<.0001)*	-0.02766(0.7337)	-0.06396(<.0001)*	-0.06554(<.0001)*	-0.3115 (0.0191)	-0.06006 (<.0001)*	-0.05836 (0.4547)	-0.06602 (<.0001)*	-0.06069 (0.4413)
b11	0.06882(<.0001)*	0.06347(<.0001)*	-0.3885(<.0001)*	0.08234(<.0001)*	0.07744(<.0001)*	-0.5438 (0.0023)	0.06354 (<.0001)*	-0.4238 (<.0001)*	0.07839 (<.0001)*	-0.4255 (<.0001)*
b12	0.09244(<.0001)*	0.08953(<.0001)*	-0.1105(0.0812)*	0.08748(<.0001)*	0.08577(<.0001)*	-0.1940 (0.0541)	0.08970 (<.0001)*	-0.1194 (0.0494)*	0.08713 (<.0001)*	-0.1309 (0.0327)*
Alpha				0.3161(<.0001)*		0.3160 (<.0001)			3.4769 (<.0001)*	

[11] *significant at 5%

Table 4.8.1 shows that for the log-linear parts, the parameter estimate of b3 (MaleStatus) was insignificant for all models unlike in the previously examined sample sizes where MaleStatus was significant for all models. In addition, the parameter estimate of b9 (Solemnisation) was insignificant only for NBRM and NBHM unlike in previously examined sample sizes where the said variable was insignificant for all models. For the logit parts, the parameter estimate of b11 (number of children) was still significant in all the proposed two-part models as in the previously examined sample sizes. Generally, there were more significant variables that explained the logit parts of the models in the 50% sample size. The dispersion parameters for all negative Binomial models were significantly different from 0, but the dispersion parameter for NBHM was at par with those of the 10% and 20% samples. One may remark that the large values of alpha noticed under the 30% and 40% samples were not associated with the increase in the sample size.

Table 4.8.2 Within-Sample Model Comparison and Selection for 50% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(13.50414)>1,96	ZIP
			AIC	PRM(83309)>ZIP(80785)	ZIP
			BIC	PRM(83403)>ZIP(80974)	ZIP
			Mc Fadden R^2	PRM(0,638604)>ZIP(0,621907)	PRM
2	ZIP	NBRM	VUONG	V(39.7795)>1,96	NBRM
			AIC	ZIP(80785)>NBRM(67384)	NBRM
			BIC	ZIP(80974)>NBRM (67486)	NBRM
			Mc Fadden R^2	ZIP(0,621907)>NBRM(0,54921)	ZIP
3	NBRM	ZINB	VUONG	V (114.7747)>1,96	ZINB
			AIC	NBRM(67384)>ZINB(67240)	ZINB
			BIC	NBRM(67486)>ZINB(67436)	ZINB
			Mc Fadden R^2	NBRM(0,54921) <ZINB(0,5503457)	ZINB
4	ZINB	PHM	VUONG	V(-42.29425)<-1,96	ZINB
			AIC	ZINB(67240)<PHM(80768)	ZINB
			BIC	ZINB(67436)<PHM(80956)	ZINB
			Mc Fadden R^2	ZINB(0,5503457) <PHM(0,621987)	PHM
5	ZINB	NBHM	AIC	ZINB(67240)>NBHM(67207)	NBHM
			BIC	ZINB(67436)>NBHM(67402)	NBHM
			Mc Fadden R^2	ZINB(0,550346) >NBHM(0,54938)	ZINB

By observing the results in Table 4.8.2, for all stages, AIC and BIC agreed in terms of which model was the best, but other criteria contradicted these statistics in some stages of the within-sample comparison. As in the preceding sample sizes, AIC and BIC are implemented in selecting the best model for the 50% sample size because these criteria agreed at each stage.

Table 4.8.3 Model Comparison Based on AIC and BIC only: 50% Sample Size[12]

MODEL	AIC	MODEL	BIC
NBHM	67207	NBHM	67402
ZINB	67240	ZINB	67436
NBRM	67384	NBRM	67486
PHM	80768	PHM	80956
ZIP	80785	ZIP	80974
PRM	83309	PRM	83403

The results in Table 4.8.3 shows that all Poisson-based models had relatively high AIC and BIC compared to Negative Binomial-based models. In general, NBHM had the lowest AIC and BIC when compared to the other five models under study, hence it was considered for the between-sample comparison phase later in the study.

[12] AIC and BIC are sorted in ascending order

Table 4.9.1 Parameter Estimates for the Models Implemented in the 60% Sample

Parameter	PRM	ZIP		NBRM	PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	0.5136(<.0001)*[13]	0.5173(<.0001)*	-4.5287(0.0138)*	0.1521(0.0001)*	0.5236(<.0001)*	-2.8050(0.0217)	0.1340(0.0010)*	-1.4453(0.0112)*
b1	-0.00252(0.0045)*	-0.00269(0.0025)*	-0.01714(0.7927)	-0.00043(0.7608)	-0.00267(0.0027)*	0.007122(0.8781)	-0.00051(0.7159)	-0.00627(0.7692)
b2	0.000963(0.2431)	0.001056(0.2007)	0.03755(0.5324)	0.001797(0.1676)	0.001043(0.2070)	0.02155(0.6211)	0.001843(0.1607)	0.005200(0.7936)
b3	-0.03227(0.0032)*	-0.03151(0.0040)*	0.2147(0.7242)	-0.06377(<.0001)*	-0.02766(0.0131)*	0.2864(0.4107)	-0.05821(0.0001)*	0.3351(0.1629)
b4	-0.06066(<.0001)*	-0.06024(<.0001)*	0.3280(0.4617)	-0.08684(<.0001)*	-0.05503(<.0001)*	0.2236(0.4839)	-0.08253(<.0001)*	0.4054(0.1333)
b5	-0.2742(<.0001)*	-0.2754(<.0001)*	-0.5388(0.5948)	-0.2187(<.0001)*	-0.2825(<.0001)*	-0.1048(0.8420)	-0.2320(<.0001)*	-0.5343(0.2062)
b6	-0.3103(<.0001)*	-0.3087(<.0001)*	0.1108(0.8531)	-0.2546(<.0001)*	-0.3191(<.0001)*	0.2315(0.5520)	-0.2644(<.0001)*	-0.5064(0.2637)
b7	0.1957(<.0001)*	0.1965(<.0001)*	-0.04976(0.8891)	0.2161(<.0001)*	0.1962(<.0001)*	-0.4121(0.1155)	0.2186(<.0001)*	0.1811(0.1196)
b8	0.4038(<.0001)*	0.4025(<.0001)*	-0.3649(0.3080)	0.4474(<.0001)*	0.4033(<.0001)*	-0.4740(0.0778)	0.4518(<.0001)*	-0.3355(0.0060)*
b9	0.01326(0.0346)	0.01325(0.0349)	0.08895(0.8560)	0.01374(0.1684)	0.01340(0.0330)	0.1088(0.7515)	0.01410(0.1615)	-0.5185(0.0004)*
b10	-0.05458(<.0001)*	-0.05415(<.0001)*	0.06910(0.7427)	-0.05535(<.0001)	-0.05432(<.0001)	0.05230(0.7314)	-0.05599(<.0001)*	-0.05943(0.4025)
b11	0.1351(<.0001)*	0.1335(<.0001)*	-1.3689(0.0021)*	0.1581(<.0001)*	0.1336(<.0001)*	-1.1921(<.0001)*	0.1594(<.0001)*	-0.4580(<.0001)*
b12	0.09713(<.0001)*	0.09699(<.0001)*	0.07458(0.6690)	0.09728(<.0001)*	0.09714(<.0001)*	0.01167(0.9241)	0.09836(<.0001)*	-0.1187(0.0346)*
Alpha				0.1151(<.0001)*			8.6566(<.0001)*	

[13] *significant at 5%

Table 4.9.1 shows that the parameter estimates of b1 (MaleOccupation) was insignificant for NBHM and b2 (FemaleOccupation) was insignificant in all models for both the log-linear and logit parts. The estimate of b9 (Solemnisation) was insignificant for all models except for the logit part of NBHM. ZINB failed to converge and did not reach second order optimality hence there were no parameter estimates for the said model. The failure to converge of ZINB may be associated with a large sample size. The dispersion parameters for all negative Binomial models were significantly different from 0. This implies that the data were over-dispersed. The dispersion parameter for NBHM is very large relative to other sample sizes.

Table 4.9.2 Within-Sample Model Comparison and Selection for 60% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(2.924332)>1,96	ZIP
			AIC	PRM(82167)>ZIP(79258)	ZIP
			BIC	PRM(82264)>ZIP(79452)	ZIP
			Mc Fadden R^2	PRM(0,68670)>ZIP(0,68607)	PRM
2	ZIP	NBRM	VUONG	V(4.92629)>1,96	NBRM
			AIC	ZIP(79258)>NBRM(74424)	NBRM
			BIC	ZIP(79452)>NBRM (74528)	NBRM
			Mc Fadden R^2	ZIP(0,68607)>NBRM(0,57710)	ZIP
3	NBRM	ZINB	AIC	NBRM(774424)	**Hessian matrix is full rank but has at least one negative eigenvalue. Second-order optimality condition violated.**
			BIC	NBRM(74528)	
			Mc Fadden R^2	NBRM(0,57710)	
4	NBRM	PHM	VUONG	V(-4.78298)< -1,96	NBRM
			AIC	NBRM(74424)<PHM(79243)	NBRM
			BIC	NBRM(74528)<PHM(79436)	NBRM
			Mc Fadden R^2	NBRM(0,57710) <PHM(0,68614)	PHM
5	NBRM	NBHM	LRT	LRT=911; Df=12; P-Value <0.001	NBHM
			AIC	NBRM(74424)<NBHM(73539)	NBHM
			BIC	NBRM(74528)>NBHM(73740)	NBHM
			Mc Fadden R^2	NBRM(0,57710)<NBHM(0,58866)	NBHM

The results in Table 4.9.2 are similar to the ones in Table 4.4.3 in which AIC and BIC agreed in terms of which model was the best, but other statistics contradicted these selection criteria in some of the stages of the within-sample comparison. As such, AIC and BIC were implemented in selecting the best model for the 60% sample size because these selection criteria agree at each stage.

Table 4.9.3 Model Comparison Based on AIC and BIC only: 60% Sample Size[14]

MODEL	AIC	MODEL	BIC
NBHM	73539	NBHM	73740
NBRM	74424	NBRM	74528
PHM	79243	PHM	79436
ZIP	79258	ZIP	79452
PRM	82167	PRM	82264

Table 4.9.3 shows that all Poisson-based models have relatively high AIC and BIC compared to Negative Binomial-based models. In overall, NBHM has the lowest AIC and BIC compared to the other five models under study, hence it is considered for the between-sample comparison phase.

[14] AIC and BIC are sorted in ascending order

Table 4.10.1 Parameter Estimates for the Models Implemented in the 70% Sample

Parameter	PRM	ZIP		NBRM	PHM		NBHM	
	log-linear	log-linear	Logit	log-linear	log-linear	Logit	log-linear	Logit
b0	0.4928(<.0001)*[15]	-4.1734(0.0130)*	0.4967(<.0001)*	0.1227(0.0009)*	0.5024(<.0001)*	-2.5892(0.0231)*	0.1009(0.0074)*	-1.4918(0.0030)*
b1	-0.00206(0.0123)*	-0.03773(0.5375)	-0.00224(0.0064)*	-0.00008(0.9507)	-0.00222(0.0069)*	-0.00688(0.8735)	-0.00029(0.8234)	-0.02214(0.2590)
b2	0.000540(0.4779)	0.04499(0.4183)	0.000640(0.4005)	0.001484(0.2171)	0.000631(0.4079)	0.02808(0.4865)	0.001574(0.1949)	0.006388(0.7255)
b3	-0.04482(<.0001)*	0.3032(0.6197)	-0.04361(<.0001)*	-0.07628(<.0001)*	-0.04008(<.0001)*	0.2813(0.4161)	-0.07278(<.0001)*	0.2126(0.3199)
b4	-0.04828(<.0001)*	0.4815(0.1968)	-0.04785(<.0001)*	-0.08056(<.0001)*	-0.04296(0.0002)*	0.3690(0.1926)	-0.07506(<.0001)*	0.2786(0.1855)
b5	-0.2506(<.0001)*	-0.8184(0.4435)	-0.2529(<.0001)*	-0.1994(<.0001)*	-0.2596(<.0001)*	-0.2389(0.6567)	-0.2092(<.0001)*	-0.4469(0.2170)
b6	-0.3273(<.0001)*	0.1940(0.6837)	-0.3244(<.0001)*	-0.2613(<.0001)*	-0.3342(<.0001)*	0.2743(0.4295)	-0.2716(<.0001)*	-0.1968(0.5497)
b7	0.1996(<.0001)*	-0.00351(0.9913)	0.2003(<.0001)*	0.2229(<.0001)*	0.2001(<.0001)*	-0.3611(0.1328)	0.2249(<.0001)*	0.2076(0.0509)*
b8	0.4013(<.0001)*	-0.4263(0.1909)	0.3999(<.0001)*	0.4440(<.0001)*	0.4006(<.0001)*	-0.5684(0.0222)*	0.4486(<.0001)*	-0.3602(0.0013)*
b9	0.01478(0.0107)*	0.07358(0.8667)	0.01476(0.0108)*	0.01380(0.1336)	0.01502(0.0096)*	0.1672(0.5975)	0.01556(0.0942)	-0.5465(<.0001)*
b10	-0.05369(<.0001)*	-0.02120(0.9147)	-0.05346(<.0001)*	-0.05478(<.0001)*	-0.05361(<.0001)*	0.008774(0.9518)*	-0.05528(<.0001)*	-0.04564(0.4750)
b11	0.1378(<.0001)*	-1.3904(0.0011)*	0.1363(<.0001)*	0.1614(<.0001)*	0.1364(<.0001)*	-1.2530(<.0001)*	0.1626(<.0001)*	-0.4502(<.0001)*
b12	0.09681(<.0001)*	0.05153(0.7519)	0.09667(<.0001)*	0.09812(<.0001)*	0.09680(<.0001)*	-0.01869(0.8698)	0.09911(<.0001)*	-0.1281(0.0129)*
Alpha				0.1153(<.0001)*			8.6121(<.0001)*	

[15] *significant at 5%

Table 4.10.1 shows that for the log-linear parts, the parameter estimates of b1 (MaleOccupation) were insignificant for ZIP, NBRM and NBHM, whereas the estimates of b2 (FemaleOccupation) were insignificant for all models. ZINB failed to converge, hence there were no parameter estimates reported for it. The dispersion parameters for both NBRM and NBHM were significantly more than 0 indicating that the data are over-dispersed. The dispersion parameter for NBHM was quite large as in the 40% and 60% samples.

Table 4.10.2 Within-Sample Model Comparison and Selection for 70% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(2.957886)>1.96	ZIP
			AIC	PRM(91181)>ZIP(90958)	ZIP
			BIC	PRM(91280)>ZIP(91155)	ZIP
			Mc Fadden R^2	PRM(0.68507)<ZIP(0.68437)	ZIP
2	ZIP	NBRM	VUONG	V(4.92629)>1.96	NBRM
			AIC	ZIP(90958)>NBRM(87316)	NBRM
			BIC	ZIP(91155)>NBRM(87423)	NBRM
			Mc Fadden R^2	ZIP(0.68437)>NBRM(0.57518)	ZIP
3	NBRM	ZINB	AIC	NBRM(87316)	**Hessian matrix is full rank but has at least one negative eigenvalue. Second-order optimality condition violated.**
			BIC	NBRM(87423)	
			Mc Fadden R^2	NBRM(0.57518)	
4	NBRM	PHM	VUONG	V(-5.09293)<-1.96	NBRM
			AIC	NBRM(87316)<PHM(90938)	NBRM
			BIC	NBRM(87423)<PHM(91135)	NBRM
			Mc Fadden R^2	NBRM(0.57518) <PHM(0.68444)	PHM
5	NBRM	NBHM	LRT	LRT=1065; Df=12; (P-Value <0.001)	NBHM
			AIC	NBRM(87316)<NBHM(86277)	NBRM
			BIC	NBRM(87423)<NBHM(86482)	NBRM
			Mc Fadden R^2	NBRM(0.57518)<NBHM(0.586296)	NBHM

It can be noticed from Table 4.10.2 that for all stages, AIC and BIC agree in terms of which model is the best, but other comparison statistics contradicted these model selection criteria. These results are similar to the ones in the preceding sample sizes, hence AIC and BIC were also used in selecting the best model for the 70% sample size.

Table 4.10.3 Model Comparison Based on AIC and BIC only: 70% Sample Size[16]

MODEL	AIC	MODEL	BIC
NBHM	86277	NBHM	86482
NBRM	87316	NBRM	87423
PHM	90938	PHM	91135
ZIP	90958	ZIP	91155
PRM	91181	PRM	91280

Table 4.10.3 shows that all Poisson-based models have relatively high AIC and BIC compared to Negative Binomial-based models, but as in the previous sample sizes. In general, NBHM has the lowest AIC and BIC compared to the other five models under study, hence it is considered for the between-sample comparison phase.

[16] AIC and BIC are sorted in ascending order

Table 4.11.1 Parameter Estimates for the Models Implemented in the 80% Sample

Parameter	PRM	ZIP		NBRM	PHM		NBHM	
	log-linear	log-linear	Logit	log-linear	log-linear	Logit	log-linear	Logit
b0	0.5031(<.0001)*[17]	0.5075(<.0001)*	-3.8278(0.0135)*	0.1200(0.0006)*	0.5136(<.0001)*	-2.3430(0.0272)*	0.09962(0.0050)*	-1.3362(0.0035)*
b1	-0.00213(0.0055)*	-0.00234(0.0023)*	-0.05749(0.3012)	-0.00034(0.7786)	-0.00232(0.0025)*	-0.02697(0.4940)	-0.00056(0.6502)	-0.02147(0.2370)
b2	-0.00037(0.6005)	-0.00028(0.6941)	0.03240(0.5167)	0.000443(0.6944)	-0.00029(0.6808)	0.01837(0.6198)	0.000470(0.6796)	0.001837(0.9133)
b3	-0.04457(<.0001)*	-0.04297(<.0001)*	0.4605(0.4463)	-0.08152(<.0001)*	-0.03858(<.0001)*	0.5075(0.1446)	-0.07648(<.0001)*	0.1910(0.2990)
b4	-0.05413(<.0001)*	-0.05385(<.0001)*	0.4596(0.1359)	-0.08966(<.0001)*	-0.04869(<.0001)*	0.3228(0.1861)	-0.08391(<.0001)*	0.2252(0.2190)
b5	-0.2527(<.0001)*	-0.2555(<.0001)*	-1.0303(0.3415)	-0.1956(<.0001)*	-0.2638(<.0001)*	-0.5523(0.3542)	-0.2085(<.0001)*	-0.3394(0.2643)
b6	-0.3115(<.0001)*	-0.3077(<.0001)*	0.2989(0.4118)	-0.2372(<.0001)*	-0.3179(<.0001)*	0.3642(0.1955)	-0.2469(<.0001)*	-0.1294(0.6410)
b7	0.1958(<.0001)*	0.1964(<.0001)*	-0.1071(0.7229)	0.2218(<.0001)*	0.1963(<.0001)*	-0.4179(0.0596)	0.2243(<.0001)*	0.2250(0.0221)*
b8	0.4057(<.0001)*	0.4042(<.0001)*	-0.4452(0.1505)	0.4460(<.0001)*	0.4048(<.0001)*	-0.6186(0.0073)	0.4501(<.0001)*	-0.4092(<.0001)*
b9	0.01323(0.0146)*	0.01323(0.0147)*	0.09368(0.8145)	0.01437(0.0970)	0.01351(0.0128)*	0.1970(0.5012)	0.01573(0.0721)	-0.6013(<.0001)*
b10	-0.05724(<.0001)*	-0.05701(<.0001)*	0.005108(0.9769)	-0.05718(<.0001)*	-0.05709(<.0001)*	0.06580(0.6066)	-0.05780(<.0001)*	-0.06956(0.2356)
b11	0.1385(<.0001)*	0.1369(<.0001)*	-1.3927(0.0004)*	0.1620(<.0001)*	0.1370(<.0001)*	-1.3009(<.0001)	0.1632(<.0001)*	-0.4523(<.0001)*
b12	0.09652(<.0001)*	0.09648(<.0001)*	0.1183(0.4315)	0.09801(<.0001)*	0.09665(<.0001)*	0.02857(0.7859)	0.09917(<.0001)*	-0.1077(0.0235)*
Alpha				0.1174(<.0001)*			8.4486(<.0001)*	

[17] *significant at 5%

Table 4.11.1 shows that, for the log-linear parts, the parameter estimate of b1 (MaleOccupation) was insignificant for NBRM and NBHM, whereas the estimate of b2 (FeamleOccupation) was insignificant for all models in both the log-linear and logit parts. ZINB failed to converge and did not achieve second order optimality, hence there were no parameter estimates reported for ZINB. The dispersion parameters for all negative Binomial models were significantly different from 0 implying that the data were over-dispersed. The dispersion parameter for NBHM was large as in the 40%, 60% and 70% sample sizes.

Table 4.11.2 Within-Sample Model Comparison and Selection for 80% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(2.942035)>1,96	ZIP
			AIC	PRM(104421)>ZIP(104181)	ZIP
			BIC	PRM(104521)>ZIP(104382)	ZIP
			Mc Fadden R^2	PRM(0,68433)>ZIP(0,68356)	PRM
2	ZIP	NBRM	VUONG	V(6.36931)>1,96	NBRM
			AIC	ZIP((104181)>NBRM(99832)	NBRM
			BIC	ZIP(104382)>NBRM(99940)	NBRM
			Mc Fadden R^2	ZIP(,68356)<NBRM(0,57467)	ZIP
3	NBRM	ZINB	AIC	NBRM(99832)	**The final Hessian matrix is full rank but has at least one negative eigenvalue. Second-order Optimality condition violated.**
			BIC	NBRM(99940)	
			Mc Fadden R^2	NBRM(0,57467)	
4	NBRM	PHM	VUONG	V(-6.20141)<-1,96	NBRM
			AIC	NBRM(99832)<PHM(104154)	NBRM
			BIC	NBRM(99940)<PHM(104354)	NBRM
			Mc Fadden R^2	NBRM(0,57467) <PHM(0,68364)	PHM
5	NBRM	NBHM	LRT	LRT=1188; Df=12; (P-Value <0.001)	NBHM
			AIC	NBRM(99832)<NBHM(98670)	NBRM
			BIC	NBRM(99940)<NBHM(98878)	NBHM
			Mc Fadden R^2	NBRM(0,57467) <NBHM(0,58588)	NBHM

It can be noticed from Table 4.11.2 that for all stages, AIC and BIC agree in terms of which model is the best but there were contradictions between the said selection criteria and other comparison statistics in some of the stages of as in the within-sample comparison of preceding sample sizes. As such, this study considered AIC and BIC in selecting the best model for the 80% sample size.

Table 4.11.3 Model Comparison Based on AIC and BIC only: 80% Sample Size[18]

MODEL	AIC	MODEL	BIC
NBHM	98670	NBHM	98878
NBRM	99832	NBRM	99940
PHM	104154	PHM	104354
ZIP	104181	ZIP	104382
PRM	104421	PRM	104521

Table 4.11.3 shows that all Poisson-based models had relatively high AIC and BIC compared to Negative Binomial-based models, hence the results were similar to the ones in the preceding sample sizes. In overall, NBHM has the lowest AIC and BIC compared to the other five models under study, hence it is considered for the between-sample comparison phase.

[18] AIC and BIC are sorted in ascending order

Table 4.12.1 Parameter Estimates for the Models Implemented in the 90% Sample

Parameter	PRM	ZIP		NBRM	PHM		NBHM	
	log-linear	log-linear	Logit	log-linear	log-linear	Logit	log-linear	Logit
b0	0.4834(<.0001)*[19]	0.4858(<.0001)*	-4.0030(0.0073)*	0.1064(0.0012)*	0.4911(<.0001)*	-2.4627(0.0125)*	0.08223(0.0139)*	-1.3257(0.0020)*
b1	-0.00283(<.0001)*	-0.00308(<.0001)*	-0.07593(0.1619)	-0.00105(0.3611)	-0.00307(<.0001)*	-0.03409(0.3594)	-0.00129(0.2656)	-0.02145(0.2103)
b2	-0.00046(0.4974)	-0.00033(0.6194)	0.03774(0.4336)	0.000435(0.6826)	-0.00034(0.6099)	0.02407(0.4886)	0.000508(0.6372)	-0.00303(0.8491)
b3	-0.07601(<.0001)*	-0.07407(<.0001)*	0.4352(0.4611)	-0.1088(<.0001)*	-0.07133(<.0001)*	0.4224(0.1608)	-0.1065(<.0001)*	0.2141(0.2073)
b4	-0.05545(<.0001)*	-0.05589(<.0001)*	0.5198(0.0416)	-0.09255(<.0001)*	-0.05081(<.0001)*	0.3508(0.0930)	-0.08860(<.0001)*	0.1873(0.2423)
b5	-0.1997(<.0001)*	-0.2034(<.0001)*	-1.1441(0.2847)	-0.1482(<.0001)*	-0.2084(<.0001)*	-0.4412(0.3643)	-0.1565(<.0001)*	-0.3365(0.2255)
b6	-0.3141(<.0001)*	-0.3077(<.0001)*	0.3515(0.1571)	-0.2368(<.0001)*	-0.3180(<.0001)*	0.3800(0.0695)	-0.2432(<.0001)*	-0.05171(0.8244)
b7	0.1979(<.0001)*	0.1987(<.0001)*	-0.02420(0.9316)	0.2232(<.0001)*	0.1986(<.0001)*	-0.2855(0.1627)	0.2263(<.0001)*	0.2146(0.0204)*
b8	0.4045(<.0001)*	0.4029(<.0001)*	-0.4405(0.1331)	0.4450(<.0001)*	0.4033(<.0001)*	-0.7177(0.0008)*	0.4488(<.0001)*	-0.4026(<.0001)*
b9	0.01672(0.0011)*	0.01656(0.0012)*	0.05922(0.8765)	0.01678(0.0400)	0.01688(0.0010)*	0.1930(0.4800)	0.01782(0.0309)	-0.5853(<.0001)*
b10	-0.05716(<.0001)*	-0.05705(<.0001)*	-0.04354(0.7969)	-0.05797(<.0001)*	-0.05711(<.0001)*	0.02778(0.8198)	-0.05868(<.0001)*	-0.07927(0.1566)
b11	0.1371(<.0001)*	0.1357(<.0001)*	-1.4608(0.0003)	0.1608(<.0001)*	0.1356(<.0001)*	-1.3566(<.0001)*	0.1621(<.0001)*	-0.4361(<.0001)*
b12	0.09620(<.0001)*	0.09621(<.0001)*	0.1445(0.3261)	0.09771(<.0001)*	0.09635(<.0001)*	0.04250(0.6680)	0.09889(<.0001)*	-0.1244(0.0055)*
Alpha				0.1177(<.0001)*			8.4237(<.0001)*	

[19] *significant at 5%

Table 4.12.1 shows that the parameter estimates of b2 (FeamleOccupation) were insignificant for all models in both the log-linear and logit parts. ZINB failed to converge and did not achieve second order optimality, hence there were no parameter estimates reported for ZINB. The dispersion parameters for all negative Binomial models were significantly different from 0 implying that the data were over-dispersed. The dispersion parameter for NBHM was large as in the 40%, 60%, 70% and 80% sample sizes.

Table 4.12.2 Within-Sample Model Comparison and Selection for 90% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(3.202286)>1,96	ZIP
			AIC	PRM(117481)>ZIP(117200)	ZIP
			BIC	PRM(117583)>ZIP(117403)	ZIP
			Mc Fadden R^2	PRM(0,68410)>ZIP(0,68337)	PRM
2	ZIP	NBRM	VUONG	V(6.42045)>1,96	NBRM
			AIC	ZIP(117403)>NBRM(112318)	NBRM
			BIC	ZIP(120151)>NBRM(112427)	NBRM
			Mc Fadden R^2	ZIP(0,68337)>NBRM(0,57452)	ZIP
3	NBRM	ZINB	AIC	NBRM(112318)	**The final Hessian matrix is full rank but has at least one negative eigenvalue. Second-order optimality condition violated.**
			BIC	NBRM(112427)	
			Mc Fadden R^2	NBRM(0,57452)	
4	NBRM	PHM	VUONG	V(-6.29225)<-1,96	NBRM
			AIC	NBRM(112318)<PHM(117180)	NBRM
			BIC	NBRM(112427)<PHM(117383)	NBRM
			Mc Fadden R^2	NBRM(0,57452) <PHM(0,68342)	PHM
5	NBRM	NBHM	LRT	LRT=1311; Df=12; (P-Value <0.001)	NBHM
			AIC	NBRM(112318)<NBHM(111033)	NBRM
			BIC	NBRM(112427)<NBHM(111245)	NBRM
			Mc Fadden R^2	NBRM(0,57452) <NBHM(0,58570)	NBHM

It can be noticed from Table 4.12.2 are similar to the ones explained in the preceding sample sizes. AIC and BIC were implemented in selecting the best model for the 90% sample size because these selection criteria agree at each stage.

Table 4.12.3 Model Comparison Based on AIC and BIC only: 90% Sample Size[20]

MODEL	AIC	MODEL	BIC
NBHM	111033	NBHM	111245
NBRM	112318	NBRM	112427
PHM	117180	PHM	117383
ZIP	117403	PRM	117583
PRM	117481	ZIP	120151

Table 4.12.3 shows that the results are similar to the ones in previous sample sizes. In general, NBHM had the lowest AIC and BIC compared to the other five models under study, hence it was considered for the between-sample comparison phase.

[20] AIC and BIC are sorted in ascending order

Table 4.13.1 Parameter Estimates for the Models Implemented in the 100% Sample

Parameter	PRM	ZIP		NBRM	PHM		NBHM	
	log-linear	log-linear	Logit	log-linear	log-linear	Logit	log-linear	Logit
b0	1.1057(<.0001)*[21]	1.1122(<.0001)*	-4.5589(<.0001)*	1.5024(<.0001)*	1.1052(<.0001)*	-4.2226(<.0001)*	1.4839(<.0001)*	0.5917(<.0001)*
b1	-0.00111(0.0130)*	-0.00121(0.0067)*	-0.04187(0.1130)	-0.00076(0.4259)	-0.00122(0.0066)*	-0.03938(0.0897)	-0.00095(0.3334)	-0.00234(0.7662)
b2	-0.00050(0.2364)	-0.00035(0.4107)	0.03736(0.1200)	-0.00020(0.8215)	-0.00035(0.4099)	0.03833(0.0705)	-0.00008(0.9347)	0.003934(0.5949)
b3	-0.1496(<.0001)*	-0.1504(<.0001)*	-0.09737(0.4276)	-0.1539(<.0001)*	-0.1504(<.0001)*	0.04042(0.7015)	-0.1581(<.0001)*	0.1624(0.0002)*
b4	-0.1605(<.0001)*	-0.1572(<.0001)*	0.5380(<.0001)*	-0.1647(<.0001)*	-0.1573(<.0001)*	0.5256(<.0001)*	-0.1649(<.0001)*	0.1658(0.0004)*
b5	-0.02492(<.0001)*	-0.02476(<.0001)*	0.01820(0.8020)	-0.02128(<.0001)*	-0.02481(<.0001)*	0.03899(0.5227)	-0.02153(<.0001)*	0.02589(0.2104)
b6	-0.02639(<.0001)*	-0.02591(<.0001)*	0.08540(0.2269)	-0.02225(<.0001)*	-0.02598(<.0001)*	0.08674(0.1450)	-0.02222(<.0001)*	0.09919(<.0001)*
b7	0.1439(<.0001)*	0.1440(<.0001)*	-0.03614(0.6682)	0.1260(<.0001)*	0.1445(<.0001)*	-0.1028(0.1714)	0.1286(<.0001)*	-0.3928(<.0001)*
b8	0.2553(<.0001)*	0.2538(<.0001)*	-0.3552(<.0001)*	0.1859(<.0001)*	0.2547(<.0001)*	-0.3822(<.0001)*	0.1875(<.0001)*	-0.5285(<.0001)*
b9	0.01182(<.0001)*	0.01132(<.0001)*	-0.1745(0.2155)	0.006958(0.2231)	0.01157(<.0001)*	-0.1069(0.3920)	0.006639(0.2590)	-0.2301(<.0001)*
b10	-0.04200(<.0001)*	-0.04200(<.0001)*	0.006026(0.9269)	-0.05228(<.0001)*	-0.04196(<.0001)*	0.01096(0.8582)	-0.05376(<.0001)*	-0.07313(0.0010)*
b11	0.08270(<.0001)*	0.08030(<.0001)*	-1.3389(<.0001)*	0.06560(<.0001)*	0.08089(<.0001)*	-1.0139(<.0001)*	0.06498(<.0001)*	-0.3701(<.0001)*
b12	0.07354(<.0001)*	0.07440(<.0001)*	0.2349(0.0024)*	0.07670(<.0001)*	0.07443(<.0001)*	0.1524(0.0210)*	0.07918(<.0001)*	-0.1310(<.0001)*
Alpha				0.2925(<.0001)*			3.2607(<.0001)*	

[21] *significant at 5%

Table 4.13.1 shows that the parameter estimates of b2 (FemaleOccupation) was insignificant for all models in both the log-linear and logit parts. ZINB failed to converge and did not achieve second order optimality, hence there were no parameter estimates reported for ZINB. The dispersion parameters for all negative Binomial models were significantly different from 0, implying that the data were over-dispersed. The dispersion parameter for NBHM was small as compared to those of the 40%, 60%, 70% and 80% sample sizes.

Table 4.13.2 Within-Sample Model Comparison and Selection for 100% Sample

STAGE	NULL	ALT	CRITERION	COMPARE STATISTICS	PREFFERED MODEL
1	PRM	ZIP	VUONG	V(3.729241)>1,96	ZIP
			AIC	PRM(317136)>ZIP(292859)	ZIP
			BIC	PRM(317248)>ZIP(293080)	ZIP
			Mc Fadden R^2	PRM(0,288888)>ZIP(0,29161)	ZIP
2	ZIP	NBRM	VUONG	V(38.7023)>1,96	NBRM
			AIC	ZIP(293962)>NBRM(252055)	NBRM
			BIC	ZIP(294072)>NBRM(252175)	NBRM
			Mc Fadden R^2	ZIP(0,288888)>NBRM(0,15081)	ZIP
3	NBRM	ZINB	AIC	NBRM(252055)	**The final Hessian matrix is full rank but has at least one negative eigenvalue. Second-order optimality condition violated.**
			BIC	NBRM(252175)	
			Mc Fadden R^2	NBRM(0,15081)	
4	NBRM	PHM	VUONG	V(-38.3925)<-1,96	NBRM
			AIC	NBRM(252055)<PHM(292778)	NBRM
			BIC	NBRM(252175)<PHM(292999)	NBRM
			Mc Fadden R^2	NBRM(0,15081) <PHM(0,28826)	PHM
5	NBRM	NBHM	LRT	LRT=2250; Df=12; (P-Value <0.001)	NBHM
			AIC	NBRM(252055)<NBHM(249831)	NBHM
			BIC	NBRM(252175)<NBHM(250063)	NBHM
			Mc Fadden R^2	NBRM(0,15081) <NBHM(0,160842)	NBHM

It can be noticed from Table 4.13.2 that for all stages, AIC and BIC agreed in terms of which model is the best but as in the previous sample sizes, the Mc Fadden R^2 contradicted AIC and

BIC in Stage 2 and Stage 4. As such, this study employed AIC and BIC in selecting the best model for the 100% sample size.

Table 4.13.3 Model Comparison Based on AIC and BIC only: 100% Sample Size[22]

MODEL	AIC	MODEL	BIC
NBHM	249831	NBHM	250063
NBRM	252055	NBRM	252175
PHM	292778	PHM	292999
ZIP	293962	ZIP	294072
PRM	317136	PRM	317248

Table 4.13.3 shows that all Poisson-based models have relatively high AIC and BIC compared to Negative Binomial-based models. NBHM generally had the lowest AIC and BIC compared to the other five models under study, hence it was considered for the between-sample comparison phase.

4.4.1 Summary of the within-sample comparison (Part A)

The results of the within-sample comparison revealed that AIC and BIC agree in terms of which model was the best, but in some instances the Mc Fadden R^2 contradicted the AIC and BIC. The Vuong's test gave inconclusive results in cases where the NBHM was compared with the other models (see Table 4.4.2, Table 4.5.2, Table 4.6.2, Table 4.7.2, Table 4.8.2, Table 4.9.2, Table 4.10.2, Table 4.11.2, Table 4.12.2 and Table 4.13.2). Based on the consistency of AIC and BIC, these model selection criteria were considered as the final comparison selection (Table 4.4.3, Table 4.5.3, Table 4.6.3, Table 4.7.3, Table 4.8.3, Table 4.9.3, Table 4.10.3, Table 4.11.3, Table 4.12.3 and Table 4.13.3). Based on AIC and BIC, the results of the study revealed that for all sample sizes, Negative Binomial-based models (NBRM, ZINB[23] and NBHM)

[22] AIC and BIC are sorted in ascending order

[23] ZINB failed to converge for sample sizes of at least 60%, hence it was not compared to the other five models in such sample sizes.

generally had smaller AIC and BIC compared to Poisson-based models (PRM, ZIP and PHM). Smaller values of AIC and BIC for Negative Binomial-based models implied that such models fit the data better than Poisson-based models for all sample sizes.

The within-sample comparison also revealed that for the 10% and 20% sample sizes, ZIP had smaller AIC and BIC than PHM which in turn had smaller AIC and BIC than PRM. On the other hand, for sample sizes of at least 30%, PHM had smaller AIC and BIC than ZIP which in turn had lower AIC and BIC than PRM. Generally, the most basic count data model, PRM, performed poorer than all other models considered in this study. The results of the within-sample comparison also revealed that for the 10% sample size, NBHM had a small AIC than the ZINB which in turn had a smaller AIC than NBRM. However, for the same sample size, that is, 10%, NBRM had a smaller BIC than NBHM, which in turn had a smaller BIC than ZINB. Based on the results of AIC and BIC, both NBHM and NBRM were considered the best for fitting the data from the 10% sample size and these models were compared with other models in the between-sample comparison phase which is presented in the next section.

The within-sample comparison also revealed that for sample sizes from 20% to 50%, NBHM had lower AIC and BIC than ZINB which in turn had lower AIC and BIC than the NBRM. As such, the NBRM was considered the best model for fitting data from sample sizes of 20% to 50% and was compared to other models in the between-sample comparison phase. It was also noticed that ZINB failed to converge for sample sizes of at least 60%, hence this study was unable to compare ZINB with the other models for such sample sizes. For sample sizes of at least 60%, NBHM had smaller AIC and BIC than NBRM, PHM, ZIP and PRM. As such, NBHM was considered the best model to fit the data from sample sizes of at least 60% and was compared to the other models in the between-sample comparison phase for such sample sizes. In general, the within-sample comparison revealed that NBHM was best suited for fitting the

data based on the AIC and BIC. NBRM was only best for fitting the data from the 10% sample size based on BIC.

4.5 Between-Sample Model Comparison and Selection (Part A)

This section compared the models that were chosen under the within-sample comparison (10% sample size to 100% sample size) based on AIC and BIC values taking into consideration the Vuong's test and LRT for over-dispersion (see Section 4.4). The MSE and MAD were used as criteria for comparing the models in Sub-section 4.5.1. Models with smaller values of MSE and MAD were considered as good (Wegner, 2007).

4.5.1 A comparison of the best models selected in the within-sample comparison phase (Part A)

Table 4.14 Comparison of the models selected from the within-sample comparison based on MAD

SAMPLE SIZE	MODEL	MAD
10%	NBHM	33.95262
10%	NBRM	33.99139
60%	NBHM	36.68664
30%	NBHM	36.94450
40%	NBHM	37.15853
90%	NBHM	37.24991
20%	NBHM	37.60020
70%	NBHM	37.62105
80%	NBHM	37.65149
50%	NBHM	43.76130
100%	NBHM	47.65459

Table 4.14 shows that NBHM for the 10% sample size had the lowest MAD value, however one could not generally conclude that NBHM performs better in smaller samples because the MAD for NBHM did not increase consistently as the sample size increased. For example, after

the 10% sample size was the 60% sample size instead of the 20% sample size which came in the 7th position.

Table 4.15 Comparison of the models selected from the within-sample comparison based on MSE

SAMPLE	MODEL	MSE
10%	NBHM	3.92219
10%	NBRM	3.92421
20%	NBHM	3.98534
30%	NBHM	4.03030
60%	NBHM	4.03201
70%	NBHM	4.04995
90%	NBHM	4.05345
80%	NBHM	4.05818
40%	NBHM	4.09004
50%	NBHM	5.10141
100%	NBHM	5.21541

Table 4.15 shows that NBHM for the 10% sample size had the lowest MSE value, however as under the MAD, one could not generally conclude that NBHM performed better in smaller samples because the MSE for NBHM did not increase consistently as the sample size increased. For example, after the 30% sample size followed the 60% sample size instead of the 40% sample size which came in the 9th position.

Table 4.16 Comparison of the models selected from the within-sample comparison based on Pearson correlation coefficient

SAMPLE	MODEL	Pearson Correlation (Actual*Predicted)
10%	NBHM	0.73504
10%	NBRM	0.73405
60%	NBHM	0.71636
20%	NBHM	0.71582
30%	NBHM	0.71577
40%	NBHM	0.71466
70%	NBHM	0.71118
90%	NBHM	0.70931
80%	NBHM	0.70843
50%	NBHM	0.60641
100%	NBHM	0.60371

Table 4.16 shows that NBHM for the 10% sample size had the highest Pearson correlation coefficient. However, one could not generally conclude that the NBHM performed better in smaller samples because the Pearson's correlation coefficient between the actual and predicted values of the DurationOfMarriage for NBHM did not increase consistently as the sample size increased. For example, after the 10% sample size followed the 60% sample size instead of the 20% sample size which came in the 4th position.

In summary, NBHM for the 10% sample size had the lowest MSE and MAD. Also, NBHM for the 10% sample size had the biggest Pearson' correlation coefficient between the actual and predicted values of the DurationOfMarriage. As such, NBHM for the 10% sample size was selected as the best model to make predictions in this study. However, although NBHM of the smallest sample size was selected as the best model, one could not generalise that NBHM performed better in smaller sample sizes because there was no consistent increase in the MSE and MAD as the sample sizes increased. Also, there was no consistent decrease in the Pearson' correlation coefficient between the actual and predicted values as the sample size increased. The test for the overall significance of the selected model is presented next.

Table 4.17 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	-14900.7661	- 6117.1487	12	17567.2348	<0.0001

Table 4.17 shows that for NBHM of the 10% sample size, the log-likelihood values for the null and full models were -14900.7661 and - 6117.1487 respectively. The chi-squared value was 2*(-6117.1487-(-14900.7661)) =17567.2348. Since there were twelve predictor variables in the full model, the degrees of freedom for the chi-squared test was 12. This yielded a p-value < 0.0001. Thus, NBHM for the 10% sample size was also statistically significant at 5% level of significance. The histogram in Figure 4.2 presents the distribution of the actual and estimated frequencies of the values of DurationOfMarriage in order to show how close predicted values were to the actual values.

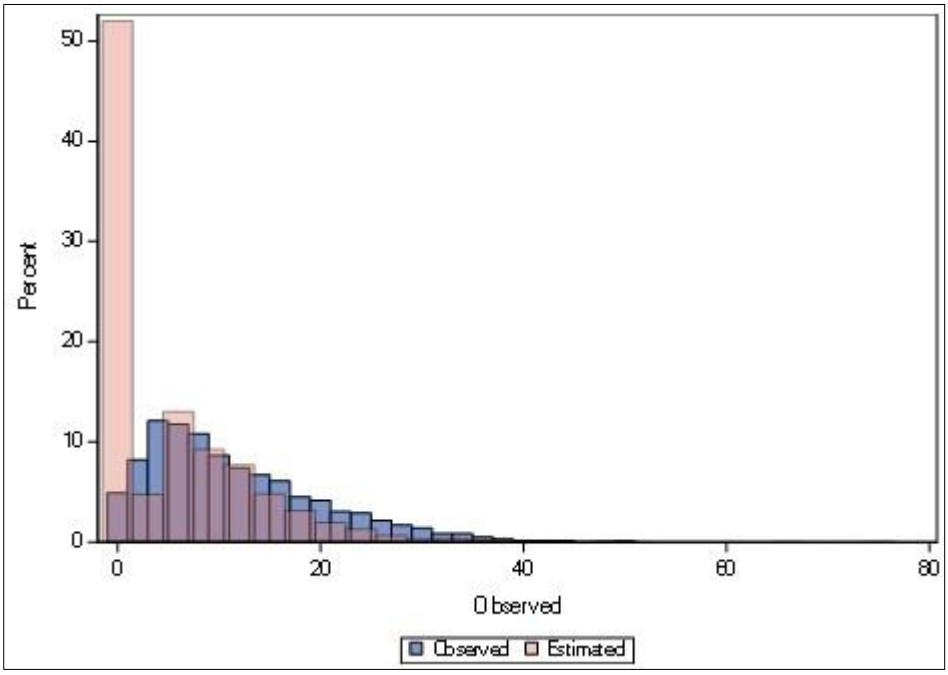

Figure 4.2 Actual versus NBHM estimated frequencies for the 10% sample size

Figure 4.2 shows that NBHM for the 10% sample size over-estimated the frequency of zero counts, but its estimates of other frequencies were closer to the observed frequencies of DurationOfMarriage even though the predicted values were slightly lower than the actual ones in general.

4.6 Key predictors of DurationOfMarriage (Part A)

Table 4.18 shows the parameter estimates for the key predictors of DurationOfMarriage from the selected model (NBHM for the 10% sample size).

Table 4.18 Predictors of the DurationOfMarriage

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
b0(Intercept)	0.06068	0.09542	2110	0.64	0.5249
b1(MaleOccupation)	0.002605	0.003476	2110	0.75	0.4537
b2(FemaleOccupation	0.001169	0.003252	2110	0.36	0.7192
b3(MaleStatus	-0.08319	0.03879	2110	-2.14	0.0321*[24]
b4(FemaleStatus	-0.1727	0.035	2110	-4.93	<.0001*
b5(MaleNoTimesMarried	-0.2775	0.05708	2110	-4.86	<.0001*
b6(FemaleNoTimesMarried)	-0.1484	0.04941	2110	-3	0.0027*
b7(MaleAge2)	0.2258	0.02086	2110	10.82	<.0001*
b8(FemaleAge2)	0.4526	0.02131	2110	21.24	<.0001*
b9(Solemnisation)	0.0271	0.02409	2110	1.12	0.2608
b10(MarriageType)	-0.03291	0.01159	2110	-2.84	0.0046*
b11(NoOfChildren)	0.1655	0.01126	2110	14.7	<.0001*
b12(CoupleRace)	0.08745	0.008914	2110	9.81	<.0001*

Table 4.18 presents predictors of DurationOfMarriage based on the parameter estimates of the selected model. One of the objectives of this study was to identify the key predictors of DurationOfMarriage. The results showed that all the 12 predictors have a role to play as far as DurationOfMarriage is concerned. However, the key predictors of DurationOfMarriage were

[24] *significant at 5%

b4 (FemaleStatus), b5 (MaleNoTimesMarried) b6 (FemaleNoTimesMarried), b7 (MaleAge2), b8 (FemaleAge2), b10 (MarriageType), b11 (NoOfChildren) and b12 (CoupleRace) which were all significant at 5% level of significance. It is worth noting that female and male occupations did not play a significant role in DurationOfMarriage. Solemnisation is another predictor of less concern in marriage life span.

4.7 Summary of data analysis and interpretation (Part A)

This chapter compared the PRM, ZIP, PHM, NBRM, ZINB and NBHM through the within-sample comparison and between-sample comparison phases. Vuong's test, LRT, Mc Fadden's R^2, AIC and BIC were used in the within-sample comparison phase. The findings revealed that, although the AIC and BIC agreed on the best models, the Mc Fadden R^2 contradicted these selection criteria in some stages of the within-sample comparison. Also, the Vuong's test gave inconclusive results for all sample sizes when other models were compared to NBHM. It was then decided that, the AIC and BIC should be used as the final comparison and selection criteria for the models under each sample size for original data analyses results.

The comparison of models based on the AIC and BIC revealed that generally, the Poisson-based models performed poorer than the Negative Binomial-based models in all sample sizes, and that, NBHM was generally the best model for fitting the data whereas PRM was the worst. The MSE, MAD and Pearson correlation coefficients between the observed and predicted values were used in the between-sample comparison phase in which all models selected as the best from the within-sample comparison were compared. The results of the between-sample comparison revealed that the NBHM for the 10% sample size fitted the DurationOfMarriage best. However, the study could not determine the effect of sample size on the efficiency of the six count data models. This because there was no consistent increase or decrease in the AIC, BIC and the Pearson correlation coefficients between the observed and predicted values of

DurationOfMarriage as the sample size increased. Note that a more detailed discussion of the results of this study is given in Chapter 5. Part B of the chapter on data analysis presents the results of the simulation study which was performed in order to investigate the efficiency of the six count data models of interest for sample sizes which were bigger than the actual data set.

Part B: Comparing the models using Monte Carlo simulated data

4.8 Distribution of DurationOfMarriage (Part B)

Table 4.19 Mean and Variance for DurationOfMarriage

n	Mean	Variance	% 0's
50 000	12.0801400	73.8061137	4.79
250 000	12.0945120	74.1883802	4.89
500 000	12.0916760	74.2376720	4.91
750 000	12.0871480	74.1700708	4.90
1000 000	12.0912310	74.3290802	4.91

Table 4.19 shows that the means and variances of the five simulated data sets did not vary across samples. It can also be noticed that the targeted count variable (DurationOfMarriage) was over-dispersed in all samples such that the variance was at least six times the mean. Figure 4.3 gave a pictorial representation of the DistributionOfMarriage and was also used for observing the possibility of zero-inflation in the target variable:

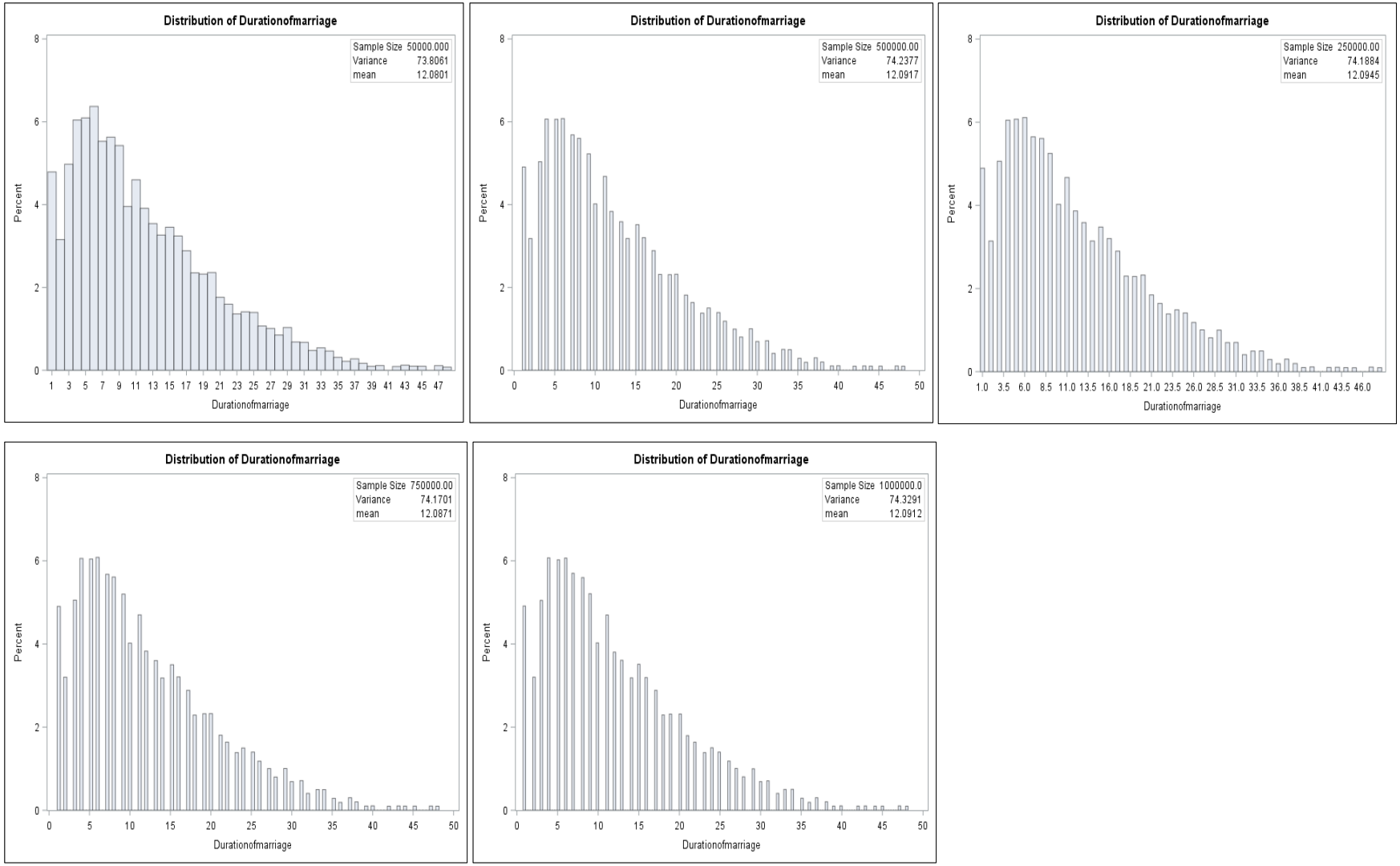

Figure 4.3 Distribution of DurationOfMarriage

Figure 4.3 confirmed the results in Table 4.19 that the distribution of DurationOfMarriage was approximately equal across all samples. Therefore, there was no bias in the dispersion parameters that may arise due to differences in the means and variances across the samples. There was no spike at zero; therefore zero inflation was unlikely (Zuur *et al*., 2009).

4.9 Diagnosis of duplicates and multicollinearity tests (Part B)

Table 4.20 Multicollinearity tests using the Variance Inflation Factor (VIF)

Variable	50 000	250 000	500 000	750 000	1000 000
MaleRace	1.00001	1.00005	1.00002	1.00001	1.00001
FemaleRace	1.00001	1.00004	1.00002	1.00002	1.00001
MaleOccupation	1.00001	1.00006	1.00004	1.00002	1.00002
FemaleOccupation	1.00002	1.00004	1.00002	1.00001	1.00001
MaleStatus	1.00001	1.00005	1.00002	1.00002	1.00001
FemaleStatus	1.00001	1.00006	1.00003	1.00001	1.00001
MaleNoTimesMarried	1.00001	1.0001	1.00003	1.00001	1.00000
FemaleNoTimesMarried	1.00000	1.00004	1.00003	1.00001	1.00001
Solemnisation	1.00001	1.00005	1.00003	1.00002	1.00001
MarriageType	1.00001	1.00006	1.00002	1.00001	1.00001
NoOfChildren	1.00001	1.00012	1.00004	1.00002	1.00001
MaleAgeTwo	1.00001	1.00003	1.00002	1.00001	1.00001
FemaleAgeTwo	1.00001	1.00005	1.00002	1.00001	1.00001

Table 4.20 shows that for all the samples, VIF<10, which implies that there is no serious multicollinearity among the independent variables. The findings in this regard are in accordance with the recommendation by Belussi and Orsi (2015).

4.10 Within-Sample Model Comparison and Selection of models (Part B)

This section presents the results of an experiment which compard the six count data models of interest using Vuong's test or LRT, AIC, BIC and Mc Fadden's R^2. These criteria were used collectively for selecting the model that fits the data best. The model with the biggest value of Mc Fadden's R^2 was preferred (Karlaftis *et al*., 2010). Moreover, the model with the smallest values of AIC and BIC was considered as having a relatively better fit (Andref, 2011 and Hilbe,

2014). The Vuong's and LRT statistics tested the hypothesis that the dispersion parameter is zero with values of the Vuong statistic less than -1.96 favouring the null model while values greater than 1.96 favouring the proposed alternative model. Inversely, the absolute values of Vuong's test smaller than 1.96 were considered to be yielding inconclusive results (Peng, 2014). The comparison started from basic models that assume equi-dispersion to more complex methods that were designed to handle under-/ over-dispersion. The significance of the parameter estimates was determined using the p-values corresponding to the Wald's test as recommended by Cameron and Trivedi (2013). Table 4.21.1 presents summary results of models' parameter estimates obtained from the Monte Carlo simulated data sets.

Table 4.21.1 Parameter Estimates for the Models Implemented on the simulated data set with n=50 000

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	2.5087 (<.0001)*[25]	2.5087 (<.0001)*	-1.8188 (0.9901)	2.5086 (<.0001)*	2.5086 (<.0001)*	-0.02339 (0.9995)	2.5086 (<.0001)*	-0.1583 (0.9978)	2.4892 (<.0001)*	-0.1627 (0.9973)
b1	-0.00138 (0.0004)*	-0.00138 (0.0004)*	-0.1658 (0.9783)	-0.00139 (0.1573)	-0.00139 (0.1565)*	-0.1537 (0.9373)	-0.00139 (0.1567)*	-0.831 (0.915)	-0.00144 (0.1682)*	-0.8481 (0.9009)
b2	-0.00112 (0.0028)*	-0.00112 (0.0028)*	-0.1872 (0.975)	-0.00113 (0.2297)	-0.00113 (0.2289)*	-0.156 (0.9345)	-0.00113 (0.2295)*	-0.8791 (0.9135)	-0.00117 (0.2412)*	-0.8937 (0.8988)
b3	-0.00098 (0.3545)	-0.00098 (0.3545)	-0.4404 (0.9883)	-0.00097 (0.7138)	-0.00097 (0.7133)*	-0.03646 (0.995)	-0.00097 (0.713)*	-0.2509 (0.9789)	-0.001 (0.7239)*	-0.2565 (0.9748)
b4	-0.00013 (0.9205)	-0.00013 (0.9192)	-0.5523 (0.9875)	-0.00013 (0.9678)	-0.00013 (0.9667)*	-0.03478 (0.9958)	-0.00013 (0.9668)*	-0.2312 (0.9798)	-0.00016 (0.9626)*	-0.238 (0.976)
b5	-0.00054 (0.4547)	-0.00054 (0.4559)	-0.2602 (0.9901)	-0.00049 (0.7834)	-0.00049 (0.7824)*	-0.03923 (0.9891)	-0.00049 (0.7828)*	-0.2965 (0.9713)	-0.00051 (0.788)*	-0.302 (0.9669)
b6	-0.00152 (0.0543)	-0.00152 (0.0543)	-0.2926 (0.9887)	-0.00151 (0.4421)	-0.00151 (0.4429)*	-0.03858 (0.992)	-0.00151 (0.4428)*	-0.2644 (0.9754)	-0.00156 (0.4545)*	-0.2694 (0.9713)
b7	-0.00012 (0.2239)	-0.00012 (0.2238)	-0.09348 (0.9707)	-0.00012 (0.6249)	-0.00012 (0.6256)*	-0.5952 (0.7871)	0.00012 (0.6244)*	-0.04209 (0.9576)	-0.00012 (0.6334)*	-0.02587 (0.9654)
b8	0.000478 (<.0001)*	0.000478 (<.0001)*	-0.1031 (0.9706)	0.000478 (0.0549)	0.000478 (0.0549)*	-0.5375 (0.8221)	0.000477 (0.0551)*	-0.2454 (0.8764)	0.000496 (0.0609)*	-0.2289 (0.8598)
b9	-0.00347 (0.1093)	-0.00347 (0.1097)	-1.2266 (0.9754)	-0.00346 (0.5216)	-0.00344 (0.523)*	-0.04065 (0.9972)	-0.00344 (0.5231)*	-0.257 (0.9853)	-0.00359 (0.532)*	-0.2653 (0.9822)
b10	0.001526 (0.1075)	0.001528 (0.107)	-1.1251 (0.9672)	0.001523 (0.5187)	0.001537 (0.515)*	-0.08247 (0.9904)	0.001534 (0.5158)*	-0.5182 (0.9645)	0.001604 (0.5227)*	-0.5334 (0.9574)
b11	0.00253 (0.0597)	0.002527 (0.06)	-0.7145 (0.9803)	0.002535 (0.4484)	0.002533 (0.4488)*	-0.04412 (0.9955)	0.002536 (0.4484)*	-0.281 (0.9761)	0.002605 (0.4639)*	-0.2897 (0.9714)
b12	-0.00272 (0.0004)*	-0.00272 (0.0004)*	-0.3309 (0.9819)	-0.00272 (0.1525)	-0.00272 (0.153)*	-0.06534 (0.9877)	-0.00272 (0.1528)*	-0.371 (0.9509)	-0.00282 (0.1632)*	-0.3838 (0.9408)
Alpha				0.4306 (<.0001)*	0.4306 (<.0001)*		0.4306 (<.0001)*		2.0523 (<.0001)*	

[25] *significant at 5%

It is worth noting that the log-linear part which models the data in Table 4.21.1 formed part of all the count data models under study. However, the logit part was only applicable to zero-inflated and hurdle models for modelling zero-inflation and deviations from equi-dispersion. All dispersion parameters (alpha) for negative binomial models were significantly greater than 0 at 5% level of significance indicating that there was significant over-dispersion in the, suggesting the PRM was not appropriate for the data set under discussion. Note that a thorough comparison between PRM and other count data models of interest is presented in Table 4.21.2. For succeeding sections, only a brief interpretation of the tables of parameter estimates was given. The reader may refer to the interpretation of Table 4.21.1 for guidance.

Table 4.21.2 Within-Sample Model Comparison and Selection for the simulated data set with n=50 000

CRITERION	NULL	ALTERNATIVE	PREFFERED MODEL
STAGE 1	**PRM**	**ZIP**	
VUONG	|-1.62753|< 1.96		Inconclusive
AIC	495949	495975	PRM
BIC	496064	496205	PRM
Mc Fadden R^2	0.000145	0.000145	Inconclusive
STAGE 2	**PRM**	**NBRM**	
LRT	(LRT=154914; df=1; P < 0.0001)		NBRM
AIC	495949	341037	NBRM
BIC	496064	341160	NBRM
Mc Fadden R^2	0.000145	0.00003	PRM
STAGE 3	**NBRM**	**ZINB**	
VUONG	412.8741>1.96		ZINB
AIC	495949	341063	ZINB
BIC	496064	341301	ZINB
Mc Fadden R^2	0.000145	0.000032	NBRM
STAGE 4	**ZINB**	**PHM**	
VUONG	163.7041>1.96		PHM
AIC	495949	341063	PHM
BIC	496064	341301	PHM
Mc Fadden R^2	0.000145	0.00004	ZINB
STAGE 5	**PHM**	**NBHM**	
AIC	339377	341063	PHM
BIC	339615	341301	PHM
Mc Fadden R^2	0.000032	0.00004	NBHM

Table 4.21.2 shows that in Stage 1, PRM was preferred over ZIP based on lowest values of AIC and BIC. The Vuong's test and Mc Fadden R^2 gave inconclusive results. Based on the results of Stage 1, PRM was compared to NBRM at Stage 2. The significant LRT as well as smallest values of AIC and BIC were in favour of NBRM and dispersion parameter (alpha) was significantly greater than zero in Table 4.21.1, therefore NBRM was more appropriate for fitting the data under study than PRM as per Stage 2. NBRM was compared to ZINB in Stage 3 where the Vuong's test, and the lowest values of AIC and BIC were in favour of ZINB. As

such, ZINB was compared to PHM in Stage 4 where the Vuong's test and the smallest values of AIC and BIC were in favour of PHM. Stage 5 compared PHM to NBHM where smallest values of AIC and BIC were in favour of PHM but NBHM slightly outperformed PHM in terms of a higher Mc Fadden's R^2. Just like in Part A, the Vuong's test gave inconclusive results for all simulated data sets ($|V|<1.96$) when comparing NBHM and other models, hence the Vuong's test was not reported for NBHM in all succeeding tables but AIC, BIC and Mc Fadden's R^2 were applied.

PHM was selected as the best model for fitting the DurationOfMarriage for the simulated data set with n=50 000 and was compared to other models later in the within-sample comparison stage. The same logic for comparing and selecting the models in the simulated data set with n=50 000 was followed for all simulated sample sizes in this study. As such, the interpretation of the steps involved in the within-sample comparison phase for the remaining simulated samples is brief. The reader may refer to this interpretation used in the within-sample comparison of the simulated data set with n=50 000 for a detailed explanation of results in the succeeding within-sample comparisons.

Table 4.22.1 Parameter Estimates for the Models Implemented on the simulated data set with n=250 000

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	2.487 (<.0001)*	-7.3232(0.8857)	2.487 (<.0001)*	2.487 (<.0001)*	2.487 (<.0001)*	-0.02455 (0.9991)	-0.1122 (0.9901)	2.487 (<.0001)*	-0.1033 (0.9959)	2.4665 (<.0001)*
b1	-0.00035 (0.0481)*	-0.01011(0.9967)	-0.00035 (0.0477)*	-0.00035 (0.4271)	-0.00035 (0.4267)	-0.1611 (0.8779)	-0.5905 (0.495)	-0.00035 (0.0475)	-0.6354 (0.7637)	-0.00036 (0.439)
b2	-0.00011 (0.4973)	-0.1794(0.9405)	-0.00011 (0.4979)	-0.00011 (0.785)	-0.00011 (0.7855)	-0.1637 (0.8626)	-0.5766 (0.5068)	-0.00011 (0.5006)	-0.6295 (0.7704)	-0.00012 (0.7901)
b3	-0.00084 (0.0759)	0.05455(0.993)	-0.00084 (0.0763)	-0.00085 (0.4758)	-0.00084 (0.4762)	-0.03838 (0.9883)	-0.1811 (0.8973)	-0.00084 (0.0766)	-0.1787 (0.9544)	-0.00087 (0.4887)
b4	0.001971 (0.0005)	-0.1424(0.9872)	0.001971 (0.0005)	0.001971 (0.1646)	0.001975 (0.1636)	-0.03659 (0.9911)	-0.1627 (0.916)	0.001974 (0.0005)*	-0.1488 (0.9649)	0.002045 (0.1755)
b5	0.000203 (0.5297)	-0.257(0.9768)	0.000205 (0.5256)	0.000212 (0.7928)	0.000212 (0.7937)	-0.041 (0.9834)	-0.1818 (0.8813)	0.000211 (0.5149)	-0.1751 (0.9481)	0.000219 (0.7994)
b6	0.00012 (0.7343)	-0.2737(0.9758)	0.00012 (0.7342)	0.000124 (0.8882)	0.000121 (0.891)	-0.04045 (0.986)	-0.1756 (0.8865)	0.000121 (0.7308)	-0.1669 (0.9521)	0.000125 (0.8945)
b7	-0.00011 (0.0131)*	-0.1372(0.9065)	-0.00011 (0.0131)*	-0.0001 1(0.3216)	-0.00011 (0.3238)	-0.6387 (0.5807)	-0.08655 (0.6054)	-0.00011 (0.0129)*	-0.1409 (0.7492)	-0.00011 (0.3372)
b8	0.000234 (<.0001)*	-0.1885(0.8964)	0.000234 (<.0001)*	0.000234 (0.0363)*	0.000235 (0.0362)*	-0.5766 (0.6377)	-0.2049 (0.4206)	0.000235 (<.0001)*	-0.2942 (0.6835)	0.000243 (0.0413)
b9	0.003184 (0.001)*	-0.6268(0.9647)	0.003184 (0.001)*	0.003171 (0.1903)	0.003162 (0.1916)	-0.0427 (0.9937)	-0.1929 (0.9321)	0.003191 (0.001)*	-0.176 (0.9725)	0.003274 (0.204)
b10	0.00048 (0.2579)	0.103(0.9843)	0.000479 (0.2584)	0.000475 (0.654)	0.00048 (0.6511)	-0.08666 (0.9785)	-0.3772 (0.8114)	0.000482 (0.2557)*	-0.3465 (0.9193)	0.000499 (0.6586)
b11	-0.00054 (0.3696)	-0.2057(0.9819)	-0.00054 (0.3711)	-0.00053 (0.724)	-0.00053 (0.7252)	-0.04636 (0.9907)	-0.2074 (0.8916)	-0.00054 (0.3678)	-0.1897 (0.9555)	-0.00054 (0.7328)
b12	-0.00081 (0.0175)	-0.4018(0.9492)	-0.00081 (0.0174)*	-0.00081 (0.3399)	-0.00082 (0.3386)	-0.06855 (0.9717)	-0.2894 (0.7624)	-0.00081 (0.0179)*	-0.2773 (0.8984)	-0.00084 (0.3519)
Alpha				0.4339 (<.0001)*	0.4339 (<.0001)*					2.0329 (<.0001)*

Table 4.22.1 shows both significant and insignificant parameters. All dispersion parameters (alpha) for negative binomial models were significantly greater than 0 at 5% level of significance indicating that there was significant over-dispersion in the DurationOfMarriage, hence implying that PRM may not be appropriate for the data set under discussion. However, a thorough comparison between PRM and other count data models of interest is presented in Table 4.22.2.

Table 4.22.2 Within-Sample Model Comparison and Selection for the simulated data set with n=250 000

CRITERION	NULL	ALTERNATIVE	PREFFERED MODEL
STAGE 1	**PRM**	**ZIP**	
VUONG	50.24343>1.96		ZIP
AIC	2490000	2490000	Inconclusive[26]
BIC	2490000	2490000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive[27]
STAGE 2	**ZIP**	**NBRM**	
VUONG	358.185>1.96		NBRM
AIC	2490000	1710000	NBRM
BIC	2490000	1710000	NBRM
Mc Fadden R^2	0	0	Inconclusive
STAGE 3	**NBRM**	**ZINB**	
VUONG	920.4611>1.96		ZINB
AIC	1710000	1710000	Inconclusive
BIC	1710000	1710000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive
STAGE 4	**ZINB**	**PHM**	
VUONG	369.4155>1.96		PHM
AIC	1710000	2490000	ZINB
BIC	1710000	2490000	ZINB
Mc Fadden R^2	0	0	Inconclusive
STAGE 5	**ZINB**	**NBHM**	
AIC	1710000	1710000	Inconclusive
BIC	1710000	1710000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive

ZIP was selected as more appropriate than PRM in Stage 1 based on the Vuong's test. Other comparison statistics yielded inconclusive results. NBRM was preferred over ZIP at Stage 2

[26] For sample sizes of at least 250 000, AIC and BIC values of Poisson-based models (PRM, ZIP and PHM) are equal, so are the AIC and BIC values of Negative Binomial-based models (NBRM, ZINB and NBHM).As such, AIC and BIC gives inconclusive results when comparing models of the same distribution (either Poisson models or Negative Binomial models).

[27] For sample sizes of at least 250 000, the log-likelihoods of the full- and intercept only-models are equal, hence the Mc Fadden's R^2 is zero for all models and in turn gives inconclusive results relative to comparing the models.

based on the Vuong's test, as well as the lowest values of AIC and BIC. Mc Fadden's R^2 gave inconclusive results. ZINB was preferred over NBRM in Stage 3 based on Vuong's test whereas other comparison statistics yielded inconclusive results. ZINB was chosen over PHM at Stage 4 based on the lowest values of AIC and BIC taking note that only the Voung's test suggested PHM as a good model. The Mc Fadden R^2 gave inconclusive results. All comparison criteria were inconclusive in Stage 5, hence neither ZINB nor NBHM was chosen as the better model. ZINB was therefore used in the between-sample comparison stage because it was selected as the best model at Stage 4. It is worth noting that most of the comparison statistics gave inconclusive results at Stage 3 and Stage 5, hence, the results of the within-sample comparison should be used and interpreted with caution.

Table 4.23.1 Parameter Estimates for the Models Implemented on the simulated data set with n=500 000

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	2.4886 (<.0001)*[28]	2.4886 (<.0001)*	-8.1247 (0.8534)	2.4886 (<.0001)*	2.4886 (<.0001)*	-0.01966 (0.9978)	2.4886 (<.0001)*	-0.6791 (0.9694)	-0.07762 (0.9891)	2.468 (<.0001)*
b1	0.000136 (0.2735)	0.000137 (0.2707)	-0.1357 (0.9424)	0.000135 (0.6623)	0.000137 (0.6593)	-0.1293 (0.7005)	0.000136 (0.2727)	0.2231 (0.7504)	-0.5187 (0.2948)	0.000142 (0.667)
b2	-0.00005 (0.6686)	-0.00005 (0.6723)	-0.2719 (0.8988)	-0.00005 (0.8573)	-0.00005 (0.864)	-0.1314 (0.6677)	-0.00005 (0.6549)	0.3918 (0.7227)	-0.5251 (0.3141)	-0.00005 (0.8745)
b3	-0.0009 (0.0074)	-0.0009 (0.0074)	-0.1993 (0.9761)	-0.0009 (0.2813)	-0.0009 (0.2825)	-0.03077 (0.9729)	-0.0009 (0.0073)*	-0.8614 (0.8325)	-0.1267 (0.8846)	-0.00093 (0.299)
b4	0.001087 (0.0067)	0.001086 (0.0067)	-0.3339 (0.9689)	0.001089 (0.2776)	0.001086 (0.2789)	-0.0293 (0.9789)	0.001086 (0.0067)*	-0.868 (0.8415)	-0.1131 (0.9061)	0.001121 (0.2939)
b5	0.000485 (0.0345)	0.000485 (0.0342)	0.08207 (0.9774)	0.000489 (0.3942)	0.000488 (0.3947)	-0.03278 (0.9617)	0.000485 (0.0342)*	-0.746 (0.8724)	-0.1321 (0.8535)	0.00051 (0.4032)
b6	-0.00003 (0.9133)	-0.00003 (0.9153)	-0.3199 (0.9677)	-0.00002 (0.9736)	-0.00003 (0.9646)	-0.0324 (0.9637)	-0.00003 (0.8988)	-0.7702 (0.8648)	-0.1269 (0.8655)	-0.00003 (0.9622)
b7	-0.00008 (0.0074)	-0.00008 (0.0075)	-0.04749 (0.9433)	-0.00008 (0.2891)	-0.00008 (0.2868)	-0.5387 (0.1049)	-0.00008 (0.0072)*	-0.2159 (0.4521)	-0.1362 (0.3003)	-0.00009 (0.3001)
b8	0.000167 (<.0001)*	0.000167 (<.0001)*	-0.05002 (0.9423)	0.000168 (0.0343)*	0.000167 (0.0352)*	-0.485 (0.1787)	0.000167 (<.0001)*	-0.0069 (0.9607)	-0.2461 (0.1905)	0.000173 (0.0396)*
b9	-0.00037 (0.5909)	-0.00037 (0.5947)	-0.6263 (0.9552)	-0.00037 (0.8306)	-0.00039 (0.8211)	-0.03418 (0.9851)	-0.00037 (0.5908)	-1.1123 (0.7333)	-0.1347 (0.9292)	-0.00038 (0.8349)
b10	0.000834 (0.0055)	0.000834 (0.0055)	-0.4545 (0.9539)	0.000834 (0.2672)	0.000836 (0.2662)	-0.06936 (0.9405)	0.000836 (0.0054)*	-1.9091 (0.2662)	-0.2722 (0.7652)	0.00086 (0.2827)
b11	-0.0004 (0.3398)	-0.00041 (0.3386)	-0.4602 (0.9545)	-0.0004 (0.7092)	-0.00039 (0.7121)	-0.0371 (0.9743)	-0.0004 (0.34)	-1.1087 (0.725)	-0.1483 (0.8813)	-0.00042 (0.7117)
b12	-0.00033 (0.1776)	-0.00032 (0.1807)	-0.05422 (0.9886)	-0.00033 (0.588)	-0.00033 (0.5888)	-0.05497 (0.9333)	-0.00033 (0.1764)	-1.1674 (0.6781)	-0.2154 (0.7219)	-0.00034 (0.5999)
Alpha				0.4345 (<.0001)*	0.4345 (<.0001)*		0.4345 (<.0001)*			2.0294 (<.0001)*

[28] *significant at 5%

Table 4.23.1 shows that both significant and insignificant parameters. All dispersion parameters (alpha) for negative binomial models were significantly greater than 0 at 5% level of significance indicating that there was significant over-dispersion in the DurationOfMarriage, implying that PRM was not be appropriate for the data set under discussion. However, a thorough comparison between PRM and other count data models of interest is presented in Table 4.23.2.

Table 4.23.2 Within-Sample Model Comparison and Selection for the simulated data set with n=500 000

CRITERION	NULL	ALTERNATIVE	PREFFERED MODEL
STAGE 1	**PRM**	**ZIP**	
VUONG	22.3453>1.96		ZIP
AIC	4980000	4980000	Inconclusive
BIC	4980000	4980000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive
STAGE 2	**ZIP**	**NBRM**	
VUONG	-507.275< -1.96		ZIP
AIC	4980000	3410000	NBRM
BIC	4980000	3410000	NBRM
Mc Fadden R^2	0	0	Inconclusive
STAGE 3	**NBRM**	**ZINB**	
VUONG	-1301.34< -1.96		NBRM
AIC	3410000	3410000	Inconclusive
BIC	3410000	3410000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive
STAGE 4	**ZINB**	**PHM**	
VUONG	523.1638> 1.96		PHM
AIC	3410000	4980000	ZINB
BIC	3410000	4980000	ZINB
Mc Fadden R^2	0	0	Inconclusive
STAGE 5	**ZINB**	**NBHM**	
AIC	3410000	3400000	NBHM
BIC	3410000	3400000	NBHM
Mc Fadden R^2	0	0	Inconclusive

ZIP was selected as more appropriate than PRM in Stage 1 based on the Vuong's test. Other comparison statistics led to inconclusive results. NBRM was preferred over ZIP at Stage 2 based on the lowest values of AIC and BIC taking note that only the Vuong's test was in favour of ZIP. Mc Fadden's R^2 gave inconclusive results. NBRM was preferred over ZINB in Stage 3 based on Vuong's test whereas other comparison statistics yielded inconclusive results. ZINB was chosen over PHM at Stage 4 based on the lowest values of AIC and BIC taking note that only the Voung's test suggested the PHM as a good model. The Mc Fadden R^2 gave inconclusive results. NBHM was chosen over ZINB in Stage 5 based on the lowest values of AIC and BIC. It is worth noting that most of the comparison statistics gave inconclusive results at Stage 1 and Stage 3, as such, the results of the within-sample comparison should be used and interpreted with caution.

Table 4.24.1 Parameter Estimates for the Models Implemented on the simulated data set with n=750 000

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	***log-linear***	***Logit***	***log-linear***	***log-linear***	***Logit***	***log-linear***	***Logit***	***log-linear***	***Logit***
b0	2.489 (<.0001)*	2.489 (<.0001)*	-12.9151 (0.6918)	2.489 (<.0001)*	2.489 (<.0001)*	-0.02543 (0.9985)	2.4891 (<.0001)*	-1.4588 (0.9825)	2.4684 (<.0001)*	-0.09763 (0.996)
b1	-0.00021 (0.0348)*	-0.00021 (0.0348)*	0.1002 (0.9496)	-0.00021 (0.399)	-0.00021 (0.3988)	-0.1682 (0.8496)	-0.00023 (0.0235)*	-0.5718 (0.8681)	-0.00022 (0.4145)	-0.6656 (0.771)
b2	-0.00003 (0.7711)	-0.00003 (0.7691)	0.04091 (0.9768)	-0.00003 (0.9087)	-0.00003 (0.9078)	-0.1711 (0.8439)	-0.00002 (0.7998)	-0.5691 (0.8705)	-0.00004 (0.8808)	-0.6797 (0.7844)
b3	-0.00095 (0.0005)*	-0.00095 (0.0005)*	-0.2164 (0.9658)	-0.00095 (0.165)	-0.00095 (0.1651)	-0.03974 (0.9875)	-0.00095 (0.0005)*	-0.9287 (0.9512)	-0.00099 (0.1768)	-0.1602 (0.9607)
b4	0.001371 (<.0001)*	0.001372 (<.0001)*	0.02017 (0.9964)	0.001372 (0.0944)	0.001372 (0.0944)	-0.03786 (0.9898)	0.001372 (<.0001)*	-1.1273 (0.9585)	0.001463 (0.0941)	-0.1379 (0.9689)
b5	0.000267 (0.1545)	0.000267 (0.155)	-0.1391 (0.9706)	0.000268 (0.5681)	0.000269 (0.5661)	-0.04242 (0.9849)	0.000265 (0.1579)	-0.4814 (0.9544)	0.000258 (0.6049)	-0.1616 (0.9537)
b6	-0.00047 (0.0221)*	-0.00047 (0.0222)*	0.1677 (0.9357)	-0.00047 (0.3615)	-0.00047 (0.3625)	-0.04188 (0.9812)	-0.00047 (0.0219)*	-0.5699 (0.9575)	-0.0005 (0.363)	-0.1687 (0.955)
b7	-0.00005 (0.0468)*	-0.00005 (0.0463)	-0.06971 (0.8988)	-0.00005 (0.4266)	-0.00005 (0.4274)	-0.7516 (0.5151)	-0.00005 (0.0422)*	-0.1485 (0.854)	-0.00005 (0.4274)	-0.1909 (0.706)
b8	0.000089 (0.0006)*	0.000088 (0.0006)*	-0.08232 (0.8903)	0.000089 (0.1702)	0.000088 (0.1705)	-0.6751 (0.5466)	0.000087 (0.0007)*	0.1733 (0.8261)	0.000095 (0.1686)	-0.3602 (0.6691)
b9	-0.00019 (0.7326)	-0.00019 (0.7351)	-0.7454 (0.9289)	-0.00019 (0.8911)	-0.00019 (0.8924)	-0.04423 (0.9919)	-0.00017 (0.7642)	-1.9667 (0.8788)	-0.00019 (0.8979)	-0.1729 (0.9742)
b10	0.001044 (<.0001)*	0.001044 (<.0001)*	-0.1024 (0.9793)	0.001043 (0.0898)	0.001043 (0.0897)	-0.08976 (0.9748)	0.001043 (<.0001)*	-2.0902 (0.7358)	0.001092 (0.0952)	-0.3437 (0.9185)
b11	0.000716 (0.0383)*	0.000716 (0.0382)	-0.06325 (0.9898)	0.000715 (0.4083)	0.000716 (0.4076)	-0.04797 (0.9878)	0.000722 (0.0367)*	-1.4675 (0.9089)	0.000732 (0.4265)	-0.1927 (0.9574)
b12	-0.00034 (0.0878)	-0.00034 (0.0877)	0.05831 (0.9827)	-0.00034 (0.4925)	-0.00034 (0.4937)	-0.07119 (0.9685)	-0.00034 (0.0852)	-0.7013 (0.8998)	-0.00035 (0.506)	-0.2712 (0.9037)
Alpha				0.4345 (<.0001)*	0.4345 (<.0001)*				2.0289 (<.0001)	

Table 4.24.1 shows both significant and insignificant parameters. All dispersion parameters (alpha) for negative binomial models were significantly greater than 0 at 5% level of significance indicating that there was significant over-dispersion in the. As such, PRM may not be appropriate for the data set under discussion. However, a thorough comparison between PRM and other count data models of interest is presented in Table 4.24.2.

Table 4.24.2 Within-Sample Model Comparison and Selection for the simulated data set with n=750 000

CRITERION	NULL	ALTERNATIVE	PREFFERED MODEL
STAGE 1	**PRM**	**ZIP**	
VUONG	-38.313< -1.96		PRM
AIC	7460000	7460000	Inconclusive
BIC	7460000	7460000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive
STAGE 2	**PRM**	**NBRM**	
LRT	LRT=2340000; Df=1; P-Value < 0.001		NBRM
AIC	7460000	5120000	NBRM
BIC	7460000	5120000	NBRM
Mc Fadden R^2	0	0	Inconclusive
STAGE 3	**NBRM**	**ZINB**	
VUONG	-1595.55 < -1.96		NBRM
AIC	5120000	5120000	Inconclusive
BIC	5120000	5120000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive
STAGE 4	**NBRM**	**PHM**	
VUONG	-621.921< -1.96		NBRM
AIC	5120000	7460000	NBRM
BIC	5120000	7460000	NBRM
Mc Fadden R^2	0	0	Inconclusive
STAGE 5	**NBRM**	**NBHM**	
LRT	LRT= 30000; Df=12; P-Value < 0.001		NBHM
AIC	5120000	5090000	NBHM
BIC	5120000	5090000	NBHM
Mc Fadden R^2	0	0	Inconclusive

Table 4.24.1 shows that PRM was selected as more appropriate than ZIP in Stage 1 based on the Vuong's test. Other comparison statistics led to inconclusive results. NBRM was preferred over ZIP at Stage 2 based on the Vuong's test and the lowest values of AIC and BIC. Mc Fadden's R^2 gave inconclusive results. NBRM was preferred over ZINB in Stage 3 based on Vuong's test whereas other comparison statistics led to inconclusive results. NBRM was chosen over PHM at Stage 4 based on the lowest values of AIC and BIC taking note that only the Voung's test suggest PHM as a good model. The Mc Fadden R^2 gave inconclusive results at Stage 4. NBHM was chosen over NBRM in Stage 5 based on a significant LRT and lowest values of AIC and BIC. It is worth noting that most of the comparison statistics gave inconclusive results at Stage 1 and Stage, hence the results of the within-sample comparison should be used and interpreted with caution.

Table 4.25.1 Parameter Estimates for the Models Implemented on the simulated data set with n=1000 000

Parameter	PRM	ZIP		NBRM	ZINB		PHM		NBHM	
	log-linear	*log-linear*	*Logit*	*log-linear*	*log-linear*	*Logit*	*log-linear*	*Logit*	*log-linear*	*Logit*
b0	2.4903 (<.0001)*[29]	-8.932 (0.7948)	2.4903 (<.0001)*	2.4903 (<.0001)*	2.4903 (<.0001)*	-0.02893 (0.9992)	2.4903 (<.0001)*	-6.3773 (0.8462)	2.4698 (<.0001)*	-0.07623 (0.9788)
b1	-0.00012 (0.1863)	-0.1428 (0.9227)	-0.00012 (0.1898)	-0.00012 (0.5994)	-0.00012 (0.5985)	-0.1934 (0.8877)	-0.00012 (0.1863)	-0.0269 (0.9845)	-0.00013 (0.5927)	-0.4949 (0.0338)*
b2	0.000032 (0.7052)	-0.1419 (0.9202)	0.000031 (0.7154)	0.000031 (0.8808)	0.000032 (0.8796)	-0.1968 (0.9171)	0.000033 (0.6907)	-0.2685 (0.8637)	0.00003 (0.8929)	-0.4915 (0.0407)*
b3	-0.00094 (<.0001)*	-0.1685 (0.9731)	-0.00094 (<.0001)*	-0.00094 (0.1118)	-0.00094 (0.113)	-0.0453 (0.9897)	-0.00094 (<.0001)	-0.2777 (0.9592)	-0.001 (0.1151)	-0.1236 (0.7686)
b4	0.001201 (<.0001)*	-0.2924 (0.9637)	0.0012 (<.0001)*	0.001205 (0.0901)	0.001203 (0.0904)	-0.04313 (0.9949)	0.001201 (<.0001)*	-0.369 (0.9548)	0.001239 (0.1016)	-0.111 (0.8174)
b5	0.000197 (0.2253)	-0.1553 (0.9703)	0.000198 (0.2241)	0.000198 (0.6269)	0.000197 (0.6283)	-0.04841 (0.99)	0.000197 (0.2252)	0.1555 (0.9323)	0.000206 (0.6336)	-0.1316 (0.7079)
b6	-0.00042 (0.0183)*	-0.4321 (0.9566)	-0.00042 (0.0186)*	-0.00042 (0.3479)	-0.00041 (0.3497)	-0.04776 (0.9906)	-0.00042 (0.0172)	0.1405 (0.948)	-0.00042 (0.3686)	-0.124 (0.7324)
b7	-0.00005 (0.0163)*	-0.04356 (0.9319)	-0.00005 (0.0165)*	-0.00005 (0.3372)	-0.00005 (0.3375)	-0.9794 (0.7657)	-0.00005 (0.0163)	-0.1706 (0.8157)	-0.00005 (0.353)	-0.1233 (0.0477)*
b8	0.000086 (0.0001)*	-0.04263 (0.9344)	0.000086 (0.0001)*	0.000086 (0.1258)	0.000086 (0.1256)	-0.8756 (0.7545)	0.000086 (0.0001)*	-0.03973 (0.9338)	0.000088 (0.1415)	-0.2105 (0.0126)*
b9	-0.00084 (0.0838)	-0.6241 (0.9428)	-0.00084 (0.0847)	-0.00084 (0.4893)	-0.00084 (0.4893)	-0.05035 (0.9956)	-0.00084 (0.0837)	-0.9858 (0.9096)	-0.0009 (0.4876)	-0.1331 (0.8597)
b10	0.000578 (0.0066)*	-0.3646 (0.9489)	0.000576 (0.0067)*	0.000577 (0.2786)	0.000578 (0.2779)	-0.1024 (0.9884)	0.000579 (0.0065)*	-0.6611 (0.9182)	0.000611 (0.2818)	-0.2665 (0.5569)
b11	0.00082 (0.0062)*	-0.3895 (0.9491)	0.000819 (0.0062)*	0.000817 (0.276)	0.000819 (0.2745)	-0.05464 (0.9933)	0.000817 (0.0063)*	-0.4769 (0.9365)	0.000874 (0.2735)	-0.1453 (0.767)
b12	-0.00000 (0.958)	-0.08916 (0.9765)	-0.00000 (0.9552)	-0.00000 (0.9828)	-0.00000 (0.9836)	-0.08124 (0.9824)	-0.00000 (0.9615)	-0.1507 (0.9594)	-0.00000 (0.9884)	-0.2078 (0.4817)
Alpha				0.4352 (<.0001)*	0.4352 (<.0001)*				2.0253 (<.0001)*	

[29] *significant at 5%

Table 4.25.1 shows that both significant and insignificant parameters. All dispersion parameters (alpha) for negative binomial models were significantly greater than 0 at 5% level of significance indicating that there was significant over-dispersion in the DurationOfMarriage, hence PRM may not be appropriate for the data set under discussion. However, a thorough comparison between PRM and other count data models of interest is presented in Table 4.25.2.

Table 4.25.2 Within-Sample Model Comparison and Selection for the simulated data set with n=1000 000

CRITERION	NULL	ALTERNATIVE	PREFFERED MODEL
STAGE 1	**PRM**	**ZIP**	
VUONG	-20.5698< -1.96		PRM
AIC	9960000	9960000	Inconclusive
BIC	9960000	9960000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive
STAGE 2	**PRM**	**NBRM**	
LRT	LRT=3130000; DF=1; P-VALUE < 0.001		NBRM
AIC	9960000	6830000	NBRM
BIC	9960000	6830000	NBRM
Mc Fadden R^2	0	0	Inconclusive
STAGE 3	**NBRM**	**ZINB**	
VUONG	-1840.63< -1.96		NBRM
AIC	6830000	6830000	Inconclusive
BIC	6830000	6830000	Inconclusive
Mc Fadden R^2	0	0	Inconclusive
STAGE 4	**NBRM**	**PHM**	
VUONG	-719.08< -1.96		NBRM
AIC	6830000	9960000	NBRM
BIC	6830000	9960000	NBRM
Mc Fadden R^2	0	0	Inconclusive
STAGE 5	**NBRM**	**NBHM**	
LRT	LRT=40000; DF=12; P-Value < 0.001		NBHM
AIC	6830000	6790000	NBHM
BIC	6830000	6790000	NBHM
Mc Fadden R^2	0	0	Inconclusive

Table 4.25.2 shows that PRM was found to be more appropriate than ZIP in Stage 1 based on the Vuong's test. Other comparison statistics yielded inconclusive results. NBRM was preferred over PRM at Stage 2 based on a significant LRT, as well as lowest values of AIC and BIC. Mc Fadden's R^2 yielded inconclusive results. NBRM was preferred over ZINB in Stage 3 based on Vuong's test whereas other comparison statistics gave inconclusive results. NBRM was chosen over PHM at Stage 4 based on the lowest values of AIC, BIC and the Vuong's test. The Mc Fadden R^2 gave inconclusive results. NBHM was chosen over NBRM in Stage 5 based on the lowest values of AIC and BIC as well as a significant LRT. It is worth noting that most of the comparison statistics gave inconclusive results at Stage 1 and Stage 3, hence the results of the within-sample comparison should be used and interpreted with caution. The next section compared the models selected from the within-sample comparison stage and was intended to identify the model that is generally good from the six.

4.11 Between-Sample Model Comparison and Selection of Models Using Simulated Data sets (Part B)

This section compared the models selected from the within sample comparison stage using MSE, MAD and the Pearson correlation coefficients of the actual and predicted DurationOfMarriage from the simulated data sets.

Table 4.25.3 Comparison of best models from the within-sample comparison phase (Part B)

Sample size	Selected model within a sample size	MSE	MAD	Pearson Correlation Coefficient
50 000	PHM	**73.7873**	**6.8412**	**0.01533**
250 000	ZINB	74.1846	6.8652	0.00688
500 000	NBHM	326.3338	15.4877	0.00464
750 000	NBHM	326.1062	15.4828	0.00423
1 000 000	NBHM	326.4387	15.4876	0.00363

Table 4.25.3 showed that PHM for the simulated data set with n=50 000 had the lowest MSE and MAD values as well as the highest value of Pearson correlation coefficient between the

actual and predicted DurationOfMarriage. As such, PHM for the sample size of n=50 000 was selected as the best from all the five being compared. It can also be noticed that ZINB for the simulated data set with n=250 000 had the second lowest values of MSE and MAD as well as the second highest value of Pearson correlation coefficient between the actual and predicted DurationOfMarriage. All NBHM models were less efficient than PHM and ZINB in Table 4.25. This is because they had bigger values of MSE and MAD as well as smaller values of the Pearson correlation coefficient between the actual and predicted DurationOfMarriage than PHM and ZINB. The results of the simulation study differed from the actual data case in Part A in which NBHM for the 10% sample size was selected. Further discussion of the results of the simulation part of this study is given in Part B of Chapter 5. The next chapter presents the discussion of results and recommendations drawn from this study.

Chapter 5

Conclusions and Recommendations

Part A: The actual data case

5.1 Introduction

This chapter reflects on the findings of the current study and its implications for further research. Conclusions drawn are in relation to the objectives listed in Chapter 1. The findings of this study are compared to those of previous studies noting similarities and differences. The chapter also suggests recommendations for further research on marriage life span as well as count data models in general.

The study was focused on performing an experiment on the Poisson and Negative Binomial count data regression models and their counterparts which are the Zero-inflated and the Hurdle models. The intention was mainly to explore their performance on different sample sizes. The study was motivated by the fact that no guideline is available with regard to a specific minimum sample size to use for count data models especially if the data exhibits under- or over-dispersion and/or zero-inflation. An experimental analysis was done on the divorce data that was collected by Statistics South Africa for the periods 2010 and 2011. A total of 22936 and 20980 divorces were recorded for the years 2010 and 2011 respectively, in South Africa.

The two data sets from 2010 and 2011 were merged and ten random samples were drawn from the merged data set in multiples of 10% until 100% (N) as summarised in Table 4.1. The findings of this study have provided an insight into the performance each of the proposed models on different sample sizes when the count response variable exhibits over-dispersion. A clearer picture concerning the effect of each independent variable on the DurationOfMarriage has also been revealed by the results in Chapter 4.

5.2 Conclusions

This section highlights the most important findings with reference to the objectives. It aims to confirm whether the objectives of this study have been attained and if not, measures that could have hindered the achievement of the objectives are highlighted in the section about the evaluation of the study.

Objective 1

To explore the efficiency of count data models under different sample sizes.

Conclusion 1

None of the previous studies interrogated in this study focused on achieving the above-mentioned objective. As such, Objective 1 was the main focus of this current study. In order to achieve this objective, the study merged the two data sets collected in 2010 and 2011 on marriages and divorces by Statistics South Africa. Ten random samples were drawn from the merged data set for 2010 and 2011 ($N = 43916$) in multiples of 10% until 100% (N). The biggest sample (100%) comprises the 43916 observations from the merged data set. By so doing it became easier for the researcher to explore the performance of the six models each, under different sample sizes. It should be noted that according to literature, there is no guideline available concerning appropriate sample size when it comes to the application of count data models.

Also note that there is no theoretical reason for choosing the sample size of 10% increments. The descriptive analysis of data revealed that the variance of the DurationOfMarriage is about seven times the corresponding mean which is an indication of possible over-dispersion in the data. This distribution of the DurationOfMarriage was further confirmed with histograms which also revealed no spikes at zero, implying that the proposed models may not be affected

by zero inflation (Zuur *et al.*, 2009). Although the histograms suggested that zero-inflation may not be problematic, this study applied zero-inflated models for comparison purposes and none on these models were selected as the best in the within-sample comparison phase. This implies that zero-inflation was not significant enough to be an issue for the data sets at hand.

Empirical findings relative to the within-sample comparison revealed Negative Binomial-based models are generally more efficient than Poisson-based regression for all sample sizes. In addition, the results of this study showed that for all sample sizes, NBHM outperformed all models in terms of smallest AIC values. Also, for sample sizes of at least 20%, NBHM outperformed all the other competing models in terms of the smallest BIC but NBRM had the lowest BIC for the 10% sample size. As such, for the 10% sample size, both NBHM and NBRM were selected as the most efficient models whereas for sample sizes of at least 20%, NBHM was selected as the most efficient model.

All models selected from the within-sample comparison were then compared in the between-sample comparison phase. The between-sample comparison revealed that NBHM for the 10% sample size as the most efficient model with lowest values of MAD and MSE, and highest Pearson correlation coefficient between the actual and predicted values. Although NBHM of the smallest sample size is selected as the best model, one may not generalise that NBHM performs better in smaller sample sizes because there was no consistent increase in the MSE and MAD as the sample sizes increase and there was no consistent decrease in the Pearson' correlation coefficient between the actual and predicted values of the DurationOfMarriage as the sample size increases.

Objective 2

To use the theoretical framework of count data to assist in building models and predicting marriage life span.

Conclusion 2

In order to achieve Objective 2, this study used literature for guidance on the common methods that are used in deriving count data models to assist in constructing the six count data models under study. The trends and differences between the methods used in the previous studies were noted and collated in the chapter on literature review. Some methods used in this study were adopted from the previous studies. On the other hand, the gaps in the previous studies such as the effect of sample size variations on the efficiency of the models and the use of few model comparison criteria which may bias the results, were identified and this study attempted to fill such.

Literature on the implementation of the NLMIXED procedure of SAS® informed this study on how to program the six count data models of interest. The study successfully utilised the theoretical framework in building the proposed count data models. It was found that variables such as female occupation and male status were insignificant in most of the proposed models. The most preferred model that was chosen for this study, that is, NBHM for the 10% sample size, may be used in predicting the DurationOfMarriage but should be used with caution since it overestimates the zero counts and slightly under-estimates the other counts. Since the theory of count data models has provided guidance in building and comparing as well as selecting between the six count data models, one may remark that Objective 2 has been achieved.

Objective 3

To identify the key predictors of the DurationOfMarriage from the available predictor variables

Conclusion 3

It is worth noting that the predictor variables used in this study were suggested by literature on the significant determinants of the success of a marriage as discussed in Chapter 2 (Holman, 2006; Reis and Sprecher, 2009 ; Cox and Demmitt, 2013). This study has successfully managed to confirm the significance of some of the predictor variables that were used in the proposed models under study. The significant predictors of DurationOfMarriage were determined by significant parameter estimates of the selected model (NBHM for the 10% sample). The variables that were identified as significant predictors of the DurationOfMarriage are FemaleStatus, MaleNoTimesMarried, FemaleNoTimesMarried, MaleAge2, FemaleAge2, MarriageType, NoOfChildren and CoupleRace. The FemaleStatus is the female's status of marriage before getting married to the partner that she is undergoing divorce with and comprise the following categories: Bachelor or spinster, widower, divorcee and married.

MaleNoTimesMarried and FemaleNoTimesMarried are the number of years that the male and female has been married before. MaleAge2 and FemaleAge2 are the male and female ages respectively, which were divided by ten in order to better the convergence of the models under study. MarriageType is the matrimonial property system which comprise the following categories: with ante-nuptial contract, without ante-nuptial contract, in community of property, out of community of property, out of community of property (excluding accrual system) and in community of property (excluding accrual system). NoOfChildren and CoupleRace are the number of children that the couple has and the race of the couple, respectively.

Of the significant variables, MaleAge2, FemaleAge2, NoOfChildren and CoupleRace have a positive contribution towards the DurationOfMarriage while other variables are negatively related to the target variable. The significant determinants of the DurationOfMarriage may be used as niche areas when formulating measures that can sustain the marriage. By so doing, the couples and professionals can work on the contributing factors so as to minimise the chances of divorce.

Objective 4

To identify the most efficient count data model for predicting the DurationOfMarriage.

Conclusion 4

As in most previous studies reviewed in this study such as Yip and Yau (2005), Rose *et al.* (2006), Hoef and Boveng (2007) and Burger *et al.* (2009), AIC, BIC, LRT, Vuong's test, MSE and correlation coefficients of the actual and predicted DurationOfMarriage were used for comparing the models and for selecting the most efficient models from the proposed six. In addition, the Mc Fadden R^2 and the MAD which are seldom used in previous studies were also implemented in this study.

The within-sample comparison phase enabled this study to successfully identify the optimal model within each sample size, whereas the between-sample comparison identified the most efficient model from different sample sizes. It was evident that the Negative Binomial-based models are generally more efficient than Poisson-based models. This study identified NBHM for the 10% sample size as the most efficient model for fitting the DurationOfMarriage. Objective 3 has been achieved.

Objective 5

To use the findings in formulating suggestions for future studies

Conclusion 5

The findings of this study have informed the recommendations for further research on count data models which are discussed later in this chapter under the recommendations section. As such, objective 5 was achieved.

5.3 Empirical findings and discussion

ZINB did not converge when the sample size is at least 50%. The problem of the non-convergence of ZINB is also experienced in the study by Famoye and Singh (2006) who also noted that Lambert (1992) encountered the same challenge. Famoye and Singh (*ibid*) decided to develop the zero-inflated generalised Poisson (ZIGP) regression model for modelling over-dispersed data with too many zeros as an alternative to ZINB. However, this study suggests other approaches of attempting to achieve convergence of ZINB when the sample size gets larger (refer to the section on recommendations for further research).

The Vuong's test gave inconclusive results when comparing NBHM with other models in Stage 5 of each within-sample comparison stage. As such, the results of this study should be used with caution due to a possibility of selection bias in Stage 5 of each within-sample comparison stage. This is because the number of the model comparison criteria was reduced at Stage 5 due to the inconclusive Vuong's test results which were not reported. Generally, more complex models: zero-inflated and hurdle models were found to be more efficient in modelling the DurationOfMarriage than PRM. This shows that more complex methods are more equipped to handle deviations from equi-dispersion and zero-inflation and/or deflation of the count response variable.

The results of this study concur with those of the study by Rose *et al.* (2006) in which it was found that NBHM generally outperformed the other four models that were examined in their study. However, since ZINB failed to converge for bigger sample sizes for this study, NBHM for the 10% sample was selected as the most efficient model for modelling the DurationOfMarriage but should be used with caution because it over-estimates the zero counts. Generally, Negative Binomial-based models outperform Poisson-based models in all sample sizes, hence the former should always be considered as alternatives to the latter when the equi-dispersion assumption is violated.

5.4 Contribution of the study

The main contribution of this study to literature is that it has explored the effect of sample size on the performance of count data models. Although there has been extensive research on the efficiency of count data models, there has been no focus on how efficient are count data models under various sample sizes. Most multivariate techniques have sample size requirements, therefore it was important to enquire about the significance of sample size in the efficiency of count data models, which is something that previous studies did not focus on. In addition, this study used AIC, BIC, Vuong's test and Mc Fadden's R^2 collectively but not marginally in selecting models in the within-sample comparison phase, hence minimising the selection bias by using numerous selection criteria[30]. The use of these statistics collectively has not been done in literature.

This study has managed to show that the DurationOfMarriage is count data by successfully modelling this variable using count data models which is also something that was never done in previous studies. The study revealed that the DurationOfMarriage is over-dispersed in nature and there was evidence of zero-deflation in the data set. The study has confirmed some well-

[30] NB: These criteria did not agree in terms of model selection

known significant predictors of a successful marriage from literature by determining the significant variables that describe DurationOfMarriage.

This study also filled a gap in literature by exploring the efficiency of count data models under both the actual and simulated data set and has managed to show that simulated data sets do not usually depict the real data scenarios. This study contributed a new approach to comparing count data models across different sample sizes by ensuring that the measures of under- or over- dispersion (variance and expectation) are kept approximately constant across all samples under study. By so doing, the study managed to minimise the biasness that may arise due to differences in the dispersion of data sets under study.

This study showed that there are instances where some model comparison criteria may produce inconclusive results, for example the Vuong's test for comparing NBHM and other models. As such, this study serves as guidance for future studies that they should consider more than one model comparison criteria in case some statistics fail to yield useful results. The study has also confirmed the non-convergence of ZINB as noticed in previous studies such as that of Famoye and Singh (2006). This study extended the findings of such studies by showing that this non-convergence of ZINB is related to bigger sample sizes.

The other contribution of this study is the use of two phases (within- and between-comparisons) for comparing the six count data models, which can also serve as a guide for future research. Also, this study contributes to the parties dealing with marriages and divorces by making recommendations to such parties on the significant predictors of the DurationOfMarriage and by suggesting the proper count data models that can be used in modelling such data.

5.5 Evaluation of this study

The Vuong's test for NBHM was not reported in the current study because it gave inconclusive results for all sample sizes. Another limitation encountered in this study is that ZINB did not converge in larger samples and as discussed under the "findings and discussion" section, some previous studies also experienced the same challenges.

The study used secondary data only because it was readily available hence inexpensive. The results may not be generalisable to sample sizes that fall outside the maximum sample size explored in the current study. Generalisability may also be limited to marriage data because it was the only type of data considered in this study. The findings of this study may not be generalised to other count data models that are not applied in this study. As discussed in Chapter 3, AIC and BIC were not used in the between-sample comparison because it was noticed that they tend to increase as the sample size increases hence they would bias the results when comparing models across different sample sizes. This was also evident from the results presented.

Moreover, the Vuong's test was not used for the between-sample comparison because it led to some inconclusive results for some models in the within-sample comparison phase. The Mc Fadden R^2 was also not used in the between-sample comparison phase because it gave contradicting results than the AIC and BIC in some stages of the within-sample comparison phase. Only predictor variables that were available in the data were considered in this study. The researcher was only limited to the variables available. Qualitative aspects that may contribute to the DurationOfMarriage were also not included in this study because the methods used for analysis in this study are quantitative. The next section gives suggestions of improving on the limitations encountered in this study.

5.5 Recommendations

This section gives suggestions for policy amendment and for further studies on the subject.

5.5.1 Recommendations to parties that may be affected by divorce

Marriage counsellors may need to consider variables such as the age of the couple (both male and female) and the number of times married (both male and female) when advising the couples about reducing the risk of divorce. This is because the said variables are some of the significant predictors of the DurationOfMarriage identified in this study. Companies such as insurance policy issuers may also need to incorporate the identified significant predictors of DurationOfMarriage when drawing risk management interventions where married couples are involved. The significant variables that influence the DurationOfMarriage may also be used at individual level to evaluate one's own marriage.

5.5.2 Recommendations for further research

The findings of this study diverge from those of some previous studies when examining other sample sizes that were explored in this study, therefore it is recommended that future studies may reduce bias when selecting count data models by varying sample sizes instead of using only one sample size. The comparison may be done in two phases, namely: the within-sample comparison phase and the between-sample comparison phase. 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.

Future studies may increase the number of sample sizes being compared as well as the maximum size explored. This will assist in confirming the results of this study. The intervals between the sample sizes may also be decreased or increased from 10%. Future studies on count data models may benefit from using other measures of goodness-of-fit such as Stavins and Jaffe goodness-of-fit statistics as well as other criteria that were not used in this study. The Stavins and Jaffe goodness-of-fit statistics were not used in the current study in order to avoid cumbersome results due to many criteria in the output. 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 quantile regression model as a potential alternative (Fuzi *et al.*, 2016), the Multivariate Poisson lognormal (MVPLN) (Zhan *et al.*, 2015) and the Negative Binomial-Lindley (NB-L) (Zamani and Ismail, 2010).

Imminent research on the prediction of marriage life span needs to consider count data models, especially hurdle models for predicting the DurationOfMarriage. This study only used twelve independent variables because they formed part of the data. Future studies may explore the significance of many more variables in explaining the DurationOfMarriage. Forthcoming research may also consider qualitative determinants of the DurationOfMarriage which can be used along with the significant quantitative predictors of DurationOfMarriage identified in this study. Reliability of these findings could be safeguarded by repeating the study of the same sample on a different count data set or on marriage data of different years or countries. These could also be areas for further studies.

5.5.3 Summary (Part A)

The first part of this chapter discussed the findings and recommendations drawn from this study. Each objective was revisited to confirm whether or not it was satisfied. The study managed to achieve all its objectives and identified NBHM for the 10% sample size as most

efficient method in modelling over-dispersed DurationOfMarriage. This study also deduced that Negative Binomial-based models are more efficient than Poisson-based models in all sample sizes and that generally NBHM is the most efficient model in fitting the DurationOfMarriage. This chapter also made some recommendations to organisations that deal with issues of divorce and to further research as well.

Part B: The Monte Carlo Simulated data case

This study simulated five data sets of the following sizes: 50000, 250000, 500000, 750000 and 1000000 using Monte Carlo simulations. As in the actual data case, PRM, ZIP, PHM, NBRM, ZINB and NBHM were compared in the within-sample and between-sample comparison phases using the following criteria: AIC, BIC, McFadden R^2, Vuong's test, LRT, MSE, MAD and Pearson correlation coefficient for the actual versus predicted values of DurationOfMarriage.

The use of numerous criteria when comparing the models was to minimise the selection bias. However, AIC and BIC for Poisson-based models were equal in sample sizes of at least 250000 so are the AIC and BIC for Negative Binomial-based models. As such, AIC and BIC gave inconclusive results when comparing models of the same distribution. Also, Mc Fadden R^2 for all models applied to sample sizes of at least 250000 were 0 because the log likelihoods of the full and reduced models were equal. As such, the Mc Fadden R^2 gave inconclusive results for all sample sizes of at least 250000. In addition, as in the actual data case, the Vuong's test for comparing NBHM gave inconclusive results in all sample sizes, hence it was not reported in this study. The within-sample comparison in the simulated data case relied mostly on the measures of over-dispersion namely: LRT and Vuong's test, hence the results of the simulated data case should be used with caution and may not be generalisable to all situations of count data models.

The between-sample comparison of the simulated data case revealed that PHM for the 10% sample size is the most efficient count data model for fitting the simulated DurationOfMarriage. The selected model had the lowest MSE and MAD as well as then highest Pearson correlation coefficient for the actual versus the predicted DurationOfMarriage. It is worth noting that, unlike in the actual data case, ZINB managed to converge in simulated data sets even when the sample sizes were very large. This might imply that the results of simulated data sets are not as realistic as the actual data sets. This argument arises from the fact that other authors who used actual data have experienced the problem of non-convergence of ZINB (Famoye and Singh, 2006) as in this current study.

The results of the simulated study may therefore not be generalisable to the real life applications of count data models. This study chooses to base its recommendations on the actual data results because most of the comparison criteria in the simulated data sets gave inconclusive results. Hence, the results of the simulated data case may not be reliable due to the possibility of selection bias.

References

1. **Agresti, A.**, 2015. *Foundations of Linear and Generalized Linear Models*, New Jersey: Wiley.
2. **Almasi, A., Eshraghian, M.R., Moghimbeigi, A., Rahimi, A., Mohammad, K. and Fallahigilan, S.,** 2016. Multilevel zero-inflated Generalized Poisson regression modeling for dispersed correlated count data. *Statistical Methodology*, 30, pp.1-14.
3. **Andreff, W**., 2011. *Contemporary Issues in Sports Economics: Participation and Professional Team Sports*, Cheltonham: Edward Elgar Publishing.
4. **Bajpai, N**., 2009. *Business Statistics*, Delhi*:* Pearson Education.
5. **Belussi, F. and Orsi, L.**, 2015. *Innovation, Alliances, and Networks in High-Tech Environments*, New York: Routledge.
6. **Berk, K. and Carey, P.**, 2010. *Data analysis with Microsoft Excel,* Australia: London.
7. **Bower, J.,** 2013. *Statistical Methods for Food Science*. Chichester, West Sussex, UK: Wiley Blackwell.
8. **Burger, M., Van Oort, F. and Linders, G.J.**, 2009. On the Specification of the Gravity Model of Trade: Zeros, Excess Zeros and Zero-Inflated Estimation. *Spatial Economic Analysis*, 4, pp.167-190.
9. **Butler, A.**, Gilbert, L. and Lewis, F., 2011. A unified approach to model selection using the likelihood ratio test, *Methods in Ecology and Evolution*, 2(2), pp. 155-162.
10. **Cameron, A.C and Trivedi, P.K.**, 2013. *Regression Analysis of Count Data*, New York: Cambridge University Press.
11. **Cowen, M. and Kattan, M.** (eds), 2009. *Encyclopedia of medical decision making, Thousand Oaks,* Calif.: SAGE Publications.
12. **Cox, F. and Demmitt, K.**, 2013. *Human Intimacy: Marriage, the Family, and Its Meaning*, Belmont, CA: Cengage Learning.

13. **Encaoua, D., Hall, B., Laisney, F. and Mairesse, J.**, 2013. *The Economics and Econometrics of Innovation*. Boston, MA: Springer US.
14. **Enders, C.**, 2010. *Applied missing data analysis*, New York: Guilford Press.
15. **Famoye, F. and Singh, K. P.**, 2006. Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4, pp. 117-130.
16. **Freese, J. and Long, J.**, 2006. *Regression models for categorical dependent variables using Stata.* College Station, Tex: StataCorp LP.
17. **Fuzi, M.F.M., Jemain, A.A. and Ismail, N.,** 2016. Bayesian quantile regression model for claim count data. Insurance: *Mathematics and Economics*, 66, pp.124-137.
18. **Greene, W., 2007.** *Functional Form and Heterogeneity in Models for Count Data*, New York*:* Now Publishers Inc.
19. **Guikema, S.D. and Goffelt, J.P.,** 2008. A flexible count data regression model for risk analysis. *Risk analysis*, 28(1), pp.213-223.
20. **Haab, T. Huang, J. C and Whitehead, J.,** 2012. *Preference Data for Environmental Valuation: Combining Revealed and Stated Approaches*, London, Taylor and Francis.
21. **Hilbe, J.M.**, 2011. *Modeling Count Data*, New York: Cambridge University Press.
22. **Hilbe, J.M.**, 2014. *Modeling Count Data*, New York: Cambridge University Press.
23. **Hoef, J. M. V. and Boveng, P. L.**, 2007. Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data? : 'Ecological Society of America'.
24. **Holman, T. B.**, 2006. *Premarital Prediction of Marital Quality or Breakup: Research, Theory, and Practice*, New York, Springer US.
25. **Jackman, S., Kleiber, C., and Zeileis, A.**, 2007. Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8).

26. **Karlaftis, M.G., Mannering, F.L. and Washington, S.P.**, 2010. *Statistical and Econometric Methods for Transportation Data Analysis*, 2nd e.d, CRC Press: Boca Raton, FL.

27. **Kiernan K, Tao J.**, 2012. Gibbs P. *Tips and Strategies for Mixed Modeling with SAS/STT® Pocedures*, Proceedings of SAS Global Forum 2012, SAS Institute Inc., Cary, NC, USA.

28. **Kleinman, K.,** MCPOWER: a flexible macro suite for generating Monte Carlo power estimates for linear, logistic, and Poisson regression models. In *SAS Global Forum 2011 Proceedings*.

29. **Li, A.**, 2013. *Handbook of SAS® DATA Step Programming*, Hoboken: CRC Press.

30. **Little, T.D.**, 2013. *The Oxford Handbook of Quantitative Methods, Vol. 2: Statistical Analysis*, New York: Oxford University Press.

31. **Loeys, T., Moerkerke, B., De Smet, O. and Buysse, A.,** 2012. The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1), pp.163-180.

32. **Matignon, R.**, 2007. *Data mining using SAS Enterprise miner*, Hoboken, New Jersey: Wiley-Inter science.

33. **Maumbe, B. M. and Okello, J.**, 2013. *Technology, Sustainability, and Rural Development in Africa*, Hershey, PA, IGI Global.

34. **Mei-Chen, H., Pavlicova, M. and Nunes, E. V.**, 2011. Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *American Journal of Drug and Alcohol Abuse*, 37, pp.367-375.

35. **Merkle, E.C. and Smithson, M.**, 2013, *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*, Boca Raton, FL: CRC Press.

36. **Moghimbeigi, A., Eshraghian, M.R., Mohammad, K. and Mcardle, B.,** 2008. Multilevel zero-inflated negative Binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, 35(10), pp.1193-1202.

37. **Morel, J. and Neerchal, N.**, 2012. *Overdispersion models in SAS*, Cary, N.C.: SAS Institute.

38. **Ozmen, I. and Famoye, F.**, 2007. Count regression models with an application to zoological\ data containing structural zeros. *Journal of Data Science*, 5, pp.491-502.

39. **Park, B.-J., Lord, D. and Hart, J. D.**, 2010. Bias properties of Bayesian statistics in finite mixture of negative Binomial regression models in crash data analysis. *Accident Analysis and Prevention*, 42, pp. 741-749.

40. **Peng, J.**, 2014. Models for Injury Count Data in the U.S. National Health Interview Survey. *JSRR*, 3(17), pp. 2286-2302.

41. **Potts, J.M. and Elith, J.**, 2006. Comparing species abundance models. *Ecological Modelling*, 199(2), pp.153-163.

42. **Reid, H.M.**, 2013. *Introduction to Statistics: Fundamental Concepts and Procedures of Data Analysis*, New Delhi: SAGE Publications.

43. **Reis, H. T. and Sprecher, S.**, 2009. *Encyclopedia of Human Relationships: Vol. 1*, Thousand Oaks, SAGE Publications.

44. **Rose, C. E., Martin, S. W., Wannemuehler, K. A. and Plikaytis, B. D.**, 2006. On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data. *Journal of Biopharmaceutical Statistics*, 16, pp. 463-481.

45. **Rosner, B.**, 2010. *Fundamentals of Biostatistics*, Boston, MA: Cengage Learning.

46. **Russell, S., Meadows, L. and Russell, R.,** 2009. *Microarray technology in practice*. Amsterdam: Elsevier/Academic Press.

47. **SAS Institute**, 2010. *SAS/STAT® 9.22 User's Guide*, Cary, NC: SAS Institute.

48. **SAS Institute**, 2011. *SAS/ETS 9.3 User's Guide*, Cary: SAS Institute.

49. **SAS-Institute**, 2012. *SAS/ETS 12.1 User's Guide*, Cary, NC, Sas Institute.

50. **Spriensma, A.S., Hajos, T.R., de Boer, M.R., Heymans, M.W. and Twisk, J.W.,** 2013. A new approach to analyse longitudinal epidemiological data with an excess of zeros. *BMC medical research methodology*, 13(1), pp.27.

51. **Statistics South Africa**, 2014. *Marriages and divorces, 2012: Metadata/Statistics South Africa*, Pretoria: Statistics South Africa.

52. **Stokes, M., Davis, C. and Koch, G.**, 2012. *Categorical data analysis using SAS, third edition*, Cary, N.C.: SAS Institute Inc.

53. **Stroup, W.**, 2012. *Generalized Linear Mixed Models*, Hoboken: CRC Press.

54. **Tang, W., He, H. and Tu, X.**, 2012. *Applied categorical and count data analysis, Boca Raton*: CRC Press.

55. **Tilanus, E.W.**, 2008. SET, MERGE and beyond, Proceedings of SAS Global Forum 2008 Conference. Cary, NC: SA Institute Inc. Paper 167-2008.

56. **Vach, W.**, 2012. *Regression Models as a Tool in Medical Research*, Boca Raton, FL, Taylor and Francis.

57. **Wang, J., Xie, H., Fisher, J. F. and Press, H. E.**, 2011. *Multilevel Models: Applications using SAS®*, Berlin, De Gruyter.

58. **Wegner, T.**, 2007. *Applied business statistics*. Cape Town: Juta.

59. **Wicklin, R.,** 2013. *Simulating data with SAS*. Cary, N.C: SAS Institute.

60. **Wicklin, R.,** 2013. *Simulating data with SAS*. SAS Institute.

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

62. **Yan, X.**, 2009. *Linear Regression Analysis: Theory and Computing*, Singapore: World Scientific.

63. **Yip, K. C. H. and Yau, K. K. W.**, 2005. On modeling claim frequency data in general insurance with extra zeros. *Insurance Mathematics and Economics*, 36, pp. 153-163.

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

65. **Zhan, X., Aziz, H.A. and Ukkusuri, S.V.,** 2015. An efficient parallel sampling technique for Multivariate Poisson-Lognormal model: *Analytic Methods in Accident Research*, 8, pp. 45–60.

66. **Zuur, A., Ieno, E., Saveliev, A., Smith, G. and Walker, N.**, 2009. *Mixed effects models and extensions in ecology with R*. New York: Springer-Verlag.

Appendices

Appendix A

SAS CODES

Selecting Random samples from the 100%Sample data set

```
/* select a 10% sample */
proc sql;
create table ps168.10%SampleSize as
select A.* from ps168.100%SampleSize as A
 where RANUNI(4321) between .45 and .55 ;
quit;
```

VIF Test

```
PROC REG DATA=Data_name PLOTS(MAXPOINTS= 100000);
MODEL Y= X1 … Xn /vif;
run;
```

PRM

```
proc nlmixed data=data_name;
parms b0=0… bn=0;
mu= exp(b0+b1*X1+… +bn*Xn);
ll = -mu + y * log(mu) - log(fact(DurationOfMarriage)); model y ~ general(ll);
predict mu out = PRM_OUT (rename = (pred = Yhat)); run;
```

NBRM

```
proc nlmixed data=data_name;
  parms b0=0… bn=0 alpha=.1;
  xb= b0+b1*X1+… +bn*Xn;
 mu = exp(xb);
  m  = 1/alpha;
  ll = lgamma(y+m)-lgamma(y+1)-lgamma(m)  + y*log(alpha*mu)-(y+m)*log(1+alpha*mu);
  model y ~ general(ll); run;
```

ZIP

proc nlmixed data=data_name tech = dbldog;

parms a0=0… an=0 b0=0… bn=0;

eta0=a0+a1*X_1+… +an*X_n;

exp_eta0 = exp(eta0);

p0= exp_eta0/(1+exp_eta0);

etap= b0+b1*X_1+… +bn*X_n;

exp_etap= exp(etap);

if Y = 0 then ll = log(p0 + (1 - p0)*exp(-exp_etap));

else ll = log(1 - p0) + Y * etap - exp_etap - log(fact(Y));

model Ye ~ general(ll);

predict exp_etap out = ZIP_OUT1 (keep= pred Y rename = (pred = Yhat));

predict p0 out = ZIP_OUT2 (keep = pred Y rename = (pred = p0)); run;

ZINB

proc nlmixed data=volis UPDATE=DDFP;

parms a0=0… an=0 b0=0… bn=0;

ETA0= a0+a1*X_1+… +an*X_n;

exp_eta0 = exp(eta0);

p0= exp_eta0/(1+exp_eta0);

etap= b0+b1*X_1+… +bn*X_n ;

exp_etap= exp(etap);

if Y = 0 then ll = log(p0 + (1-p0)*((1/gamma(1))*(((1/alpha)/(1/alpha + exp_etap))**(1/alpha))));

else ll = log(1 - p0) + lgamma(Y + 1/alpha) - lgamma(Y + 1) - lgamma(1/alpha) + Y*log(alpha*exp_etap) - (Y + 1/alpha)*log(1+alpha*exp_etap);

model Y ~ general(ll);

predict exp_etap out = zinb_OUT1 percent (keep= pred Y rename = (pred = Yhat));/*estimated alpha:3.6659*/

predict p0 out = zinb_OUT2 (keep = pred Y rename = (pred = p0));

PHM

proc nlmixed data=data_name tech = dbldog;

parms a0=0… an=0 b0=0… bn=0;

eta0= a0+a1*X_1+… +an*X_n;

exp_eta0 = exp(eta0);

p0= exp_eta0/(1+exp_eta0);

etap= b0+b1*X_1+… +bn*X_n ;

exp_etap= exp(etap);

if y = 0 then ll = log(p0);

else ll = log(1 - p0) - exp_etap + DURATIONOFMARRIAGE * etap - log(1-exp(-exp_etap))-log(fact(y));

model y ~ general(ll);

predict exp_etap out = phm_out1 (keep= pred y rename = (pred = Yhat));

predict p0 out =phml_out2 (keep = pred y rename = (pred = p0)); run;

NBHM

proc nlmixed data=data_name;

y= ;

parms a0=0… an=0 b0=0… bn=0;

** Logit model;

eta_zip= a0+a1*X_1+… +an*X_n;

p0_zipe= 1/(1+exp(-1*eta_zip));

** Truncated count model;

eta_nb = b0+b1*X_1+… +bn*X_n ;

mean=exp(eta_nb);

p=1/(1+(1/v)*mean);

p0=log(p0_zipe);

p_else=log(1-p0_zipe)+y*log(1-p)-log(p**(-1*(v))-1)+lgamma(y+(v))- lgamma(v)-log(fact(y));

if y=0 then loglike=(p0);

else loglike=(p_else);

model y ~ general(loglike);

logtheta = log(v);

ods output ParameterEstimates=peHUNB_init1;

quit;title;

Vuong's test ZIP vs PRM

data PRM_VALUES (keep = poi_prob);

```
set PRM_OUT;
do i = 0 to 5;
poi_prob = pdf('PRM', i, Yhat);
if y = i then output;
end; run;
data ZIP_VALUES (keep = zip_prob);
merge zip_OUT1 zip_OUT2;
do i = 0 to 5;
if i = 0 then zip_prob = p0 + (1 - p0) * pdf('poisson', i, Yhat);
else zip_prob = (1 - p0) * pdf('poisson', i, Yhat);
if Y = i then output;
end; run;
data compare;
merge poi_values zip_values;
m = log(zip_prob / poi_prob);
run;
proc sql;
select
mean(m) as mbar,
std(m) as s,
sqrt(count(*))*mean(m)/std(m) as v
from
compare;
quit;
```

Vuong's test ZINB vs NBRM

```
data NBRM_VALUES (keep = nb_prob);
set nb_out_10_percent;
do i = 0 to 5;
nbrm_prob = (((1/4.2623)/(1/4.2623+Yhat))**(1/4.2623))*((Yhat/(1/4.2623+Yhat))**i)*(gamma(i+1/4.2623)/gamma(i+1)/gamma(1/4.2623));
if y = i then output; end; run;
data zinb_pred (keep = zinb_prob);
```

```
merge zinb_out1 zinb_out2;

do i = 0 to 5;

if i = 0 then zinb_prob = p0 + (1 - p0) * (((1/3.6659)/(1/3.6659+Yhat))**(1/3.6659))*((Yhat/(1/3.6659+Yhat))**i)*(gamma(i+1/3.6659)/gamma(i+1)/gamma(1/3.6659));

else zinb_prob = (1 - p0) * (((1/3.6659)/(1/3.6659+Yhat))**(1/3.6659))*((Yhat/(1/3.6659+Yhat))**i)*(gamma(i+1/3.6659)/gamma(i+1)/gamma(1/3.6659));

if  y= i then output;

end; run;

data compare;

merge nbrm_pred zinb_pred;

m = log(zinb_prob / nbrm_prob);

run;

proc sql;

select

mean(m) as mbar,

std(m) as s,

sqrt(count(*))*mean(m)/std(m) as v

from

compare; quit;
```

Vuong's test PRM vs NBRM

```
data phm_values (keep = hdl_prob);

merge phm_out1 phm_out2;

do i = 0 to 5;

if i = 0 then hdl_prob = p0;

else hdl_prob = (1 - p0)/(1-exp(-Yhat))* pdf('poisson', i, Yhat);

if y = i then output;

end; run;

data nbrm_pred (keep = nbrm_prob);

set nbrm_out;

do i = 0 to 5;
```

```
nbrm_prob = (((1/4.2623)/(1/4.2623+Yhat))**(1/4.2623))*((Yhat/(1/4.2623+Yhat))**i)*(gamma(i+1/4.2623)/gamma(i+1)/gamma(1/4.2623));

if y = i then output;

end; run;

data compare;

merge phml_values  nbrm_values;

m = log(phm_prob / nbrm_prob); run;

proc sql;

select

mean(m) as mbar,

std(m) as s,

sqrt(count(*))*mean(m)/std(m) as v

from

compare; quit;
```

Vuong's test ZINB vs PHM

```
data phm_pred (keep = phm_prob);

merge =phm_out1 phm_out2;

do i = 0 to 5;

if i = 0 then phm_prob = p0;

else phm_prob = (1 - p0) * pdf('poisson', i, Yhat);

if y = i then output;

end; run;

data zinb_values (keep = zinb_prob);

merge zinb_out1 zinb_out2;

do i = 0 to 5;

if i = 0 then zinb_prob = p0 + (1 - p0) * (((1/3.6659)/(1/3.6659+Yhat))**(1/3.6659))*((Yhat/(1/3.6659+Yhat))**i)*(gamma(i+1/3.6659)/gamma(i+1)/gamma(1/3.6659));

else zinb_prob = (1 - p0) * (((1/3.6659)/(1/3.6659+Yhat))**(1/3.6659))*((Yhat/(1/3.6659+Yhat))**i)*(gamma(i+1/3.6659)/gamma(i+1)/gamma(1/3.6659));

if y = i then output; end; run;

data compare;
```

```
merge phm_values zinb_values;
m = log(zinbvalues / phm_values); run;
proc sql;
select
mean(m) as mbar,
std(m) as s,
sqrt(count(*))*mean(m)/std(m) as v
from
compare; quit;
```

Vuong's test ZIP vs NBHM

```
data phm_values (keep = phm_prob);
merge phm_out1 phm_out2;
do i = 0 to 5;
if i = 0 then prm_prob = p0;
else prm_prob = (1 - p0) * pdf('poisson', i, Yhat);
if y = i then output;
end; run;
data zip_values (keep = zip_prob);
merge zip_out1 zip_out2;
do i = 0 to 5;
if i = 0 then zip_prob = p0 + (1 - p0) * pdf('poisson', i, Yhat);
else zip_prob = (1 - p0) * pdf('poisson', i, Yhat);
if y = i then output; end; run;
data compare;
merge phm_pred zip_pred;
m = log(zip_prob / phm_prob);
run;
proc sql;
select
mean(m) as mbar,
std(m) as s,
```

```
sqrt(count(*))*mean(m)/std(m) as v
from
compare;
quit;
```

Monte Carlo Simulation of Data

```
data dataname (keep=);
call streaminit (4321);
 array p[3] (p1 p2 p3);
do i = 1 to n;
Type = rand ("Table", of p[*]);
output; end; run;
proc format;
value …; run;
 proc freq data=dataname;
format Type Call.; t
Tables Type; run;
```