**Model selection and residual analysis for responses with excessive zero counts.**

**PhD Candidature**

**CONFIRMATION REPORT**

Submitted by:

Vytautas Duoba

Msc Mathematics

15 June 2021

Supervisors:

Dr Jonathan Godfrey

Dr Jonathan Marshall

School of Fundamental Sciences

Confirmation Report in Fulfilment of Confirmation of PhD Candidature

School of Fundamental Sciences

Massey University

**Summary**

Count models involve numerical responses which are restricted to non-negative integers, typically in the set . When the majority of the response values involve small integers, e.g. K is 30 or less, the response distribution is often modelled as Poisson, negative binomial, geometric, bernouilli or binomial. Many other distributions (usually involving more parameters or more complex functional forms) are used but those just mentioned are probably the most frequently used, owing to their long history and relatively simple but effective refinement over decades of use and study. The refinements include accommodating excessive zero counts, truncated data and over-dispersion.

Examples of count data include: monthly counts of traffic accidents at a particular location, number of physician visits in a year, number of different bird species in a park, gestation months before first pediatric visit, reported count of infectious diseases in a region, and many others. Often considerable inventiveness is exhibited in re-casting an analysis challenge into a form that can be addressed by count data modelling.

The modelling of the response distribution parameters via the generalized linear modelling framework extends the use of count models into the regression arena. This enables incorporation of the richness and complexities of non-linear relationships, random, fixed and mixed effects, longitudinal and spatial dependencies in order to explain the conditional mean response and residual variation in a set of data.

The extension of relatively simple functional forms involving just one or two parameters into sophisticated models involving consideration of mixtures of distributions, covariates for conditional means and random effects for correlated responses adds considerable complexity to assessing model appropriateness. By adding functional complexity to the modelling process plus more and more covariates, the data fit can be improved almost at will. But perhaps the world and its data are complex and only a complex model will suffice to represent reality? Can a statistician know when to stop adding complexity to a model?

The objective of the research to be undertaken for this PhD is to identify a methodology and set of statistical tools that are effective in identifying appropriate and effective models for arbitrary count data, even when the count model is complex such as an autoregressive and spatial one. The objective includes identifying when count models are and are not appropriate and identifying the point at which complex models only add inconsequential marginal improvements over a simpler model. Testing of the fulfilment of the objective will include analyses of some count datasets of interest in the NZ Aotearoa context, and which have not been modelled to the candidate’s knowledge. These analyses are intended to lead to publications in their own right.

**Table of Contents**

**1. Summary**

**2. Introduction**

**3. Literature Review**

Texts

Software manuals and vignettes

Journal Articles

Other

**4. Research Aims and Objectives**

Overall research aim

Research objectives

Research questions

Benefits arising from the research

**5. Methodology**

Common functional forms used in count regression modelling

Modifications to simple models to improve model fit

Regression model formulations and parameter estimation methods

Nature of residuals for count models

Using residuals to test model assumptions and fit

Identifying a successful strategy for fitting count regression models

Gaps in the process for assessing count models

**6. Data Sources**

**7. Proposed Research Schedule**

**8. Conclusions**

**9. References**

**Appendix A: Frameworks for model fitting and assessment**

**Appendix B, etc**

**Table of Figures**

**List of Tables**

**3. Literature Review**

There is an extensive corpus of published research and books that describe count models. For example, univariate count models can be traced back to as early as 1679 (Blaise Pascal and the negative binomial distribution) and 1711 (de Moivre and his work with the Poisson distribution). Poisson in 1837 studied the number of wrongful convictions in a jurisdiction, and Ladislaus Bortkiewicz produced his famous analysis of deaths by horse kicks in 1898. However, a significant development in count modelling occurred when Nelder and Wedderburn (1972) developed the theory of generalized linear models in which the Poisson distribution plays an important role. Hence most of the literature review that follows comprises material dealing with theory developed under specific broad frameworks rather than as isolated special cases. The rationale is that the main purpose of the PhD research is to identify methods, tools and processes that can be applied in a wide a range of situations.

**3.1 Texts.**

3.1.1 *Regression Analysis of Count Data*, 2nd Ed, A Colin Cameron & Pravin K Trivedi

This text appears to be the most comprehensive source of information about count models in wide generality. It covers model evaluation and testing, models for count time series, longitudinal models, random and fixed effects models, flexible functional forms, and uses many interesting examples. I intend to concentrate on this (and Gourieroux) for most of my theoretical knowledge acquisition.

3.1.2 *Regression Models for Categorical, Count, and Related Variables*, John P. Hoffman

A very practical book concentrating on applications using the stata package. Has interesting material on multivariate probit and data reduction techniques (CFA, SEM). A good text to provide relief from unrelenting theory and equations.

3.1.3 *Functional Form and Heterogeneity in Models for Count Data*, William Greene

Concentrates on Poisson and Negative binomial models with zero-inflated and hurdle additions. Covers fixed and random effects models (panel data) and the bivariate Poisson model. Excellent, mathematically clear and concise exposition of simple basic count models.

3.1.4 *Nonlinear Models for Repeated Measurement Data*, Marie Davidian & David M. Giltman

Not count-oriented but an excellent introduction to the non-linear framework and the hierarchical non-linear model. Mainly but not solely bayesian model oriented. Very readable

3.1.5 *Econometric Analysis of Panel Data*, Baltagi, Badi H

3.1.6 *Statistics and Econometric Models, Vol 1*, Christian Gourieroux, Alain Monfort

While not covering count models specifically, the value of this text is its general coverage of a wide range of statistical modelling challenges. It covers both classical and bayesian approaches. I think that together with 3.1.1 it should provide a very thorough theoretical and practical basis for more advanced research and for knowledgeable analysis work.

3.1.7 *Negative Binomial Regression*, 2nd edition, Joseph M. Hilbe, Arizona State University

Presents the varieties of the NB2 negative binomial model, a special case of which is the Poisson regression model. Also discusses Poisson Inverse Gaussian, binomial, geometric models, censoring, zero-inflation, mixed and random effects. Has useful sections on model fitting and model validation.

**3.2 Manuals, Vignettes**

3.2.1 *Temporal and Spatio-Temporal Modelling and Monitoring of Epidemic Phenomen*a, Michael Höhle, Sebastian Meyer, et alia. R ‘surveillance’ package vignette

3.2.2 *Package ‘EpiModel’*, Samuel Jenness, Steven M. Goodreau, et alia, package vignette

3.2.3 *Residual diagnostics for hierarchical (multi-level/mixed) regression models*, Florian Hartig, DHARMA package vignette

**3.3 Journal Articles**

3.3.1 *Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance*, Sebastian Meyer, Leonhard Held, Michael Höhle, Journal of Statistical Software

3.3.2 *EpiModel: An R Package for Mathematical Modelling of Infectious Disease over Networks*, Samuel M. Jenness, Steven M. Goodreau, Martina Morris, Journal of Statistical Software

3.3.3 *Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model*

Heinen, Andreas

3.3.4 *A flexible distribution class for count data*, Kimberly F. Sellers et alia

The negative binomial distribution, is unable to address data under-dispersion. This work introduces and thus considers the sum of CMP (Conway-Maxwell-Poisson) random variables to establish the flexible class of distributions (sCMP) that encompass the Poisson, geometric, Bernoulli, negative binomial, binomial, and CMP distributions as special cases.

3.3.5 *Diagnostics for Regression Models with Discrete Outcomes Using Surrogate Empirical Residual Distribution Functions*, Lu Yang

3.3.6 *Useful models for time series of counts or simply wrong ones?* Robert C. Jung · A.R. Tremayne

3.3.7 *The chi-squared goodness-of-fit test for count-data models*, Miguel Manj´on, Oscar Mart´ınez

3.3.8 *Poisson Inverse Gaussian (PIG) Model for Infectious Disease Count Data*, Vincent Moshi Ouma, et alia

The Poisson- Inverse Gaussian (PIG) model can be used to analyze count data that is highly overdispersed. A PIG model with fixed/varying dispersion parameters is fitted to two infectious disease datasets and its performance in terms of goodness-of-fit and future outbreak predictions of infectious disease is compared to that of the traditional NB model. The results are inconclusive.

**3.4 Other**

3.4.1 *Count Data: Modelling and Estimation*, Economics Working Paper, No. 2005-08, Jung, Robert, et alia

**4. Research Aims and Objectives**

**4.1 Overall research aim**

My motivation is the belief that there is an excess of count models that have been identified in the literature and that they are surplus to the requirements of competent, practising applied statisticians. The process of identifying, fitting and interpreting a count model needs to be simplified and automated to assist knowledgeable, practising statisticians.

**4.2 Research objectives**

4.2.1 Explore the range of applications of count models used in econometrics, psychology, social science, and ecology.

4.2.2 Explore how non-count data are put into a count model framework

4.2.3 identify the most useful collection of count model functional forms, enhancements and mixtures that can satisfy the majority of count modelling situations

4.2.4 List the methods and tools employed to assess the appropriateness of count models

4.2.5 Assess the marginal improvement that complex functional forms and enhancements bring to count modelling

4.2.6 identify gaps in count model fitting and assessment

**4.3 Research questions**

4.3.1 Can the different count model functional forms be identified in the main and ranked by popularity of use?

4.3.2 Does the type of application (econometric, social, psychological, market research, biological, science, etc) strongly determine the type of count model fitted?

4.3.3 Which are the most successful count model types per type of application?

4.3.4 Is there a small set of count model functional forms and regression types that appear across most application areas?

4.3.5 Is it possible to automate model selection, fitting, and diagnosis using a limited set of functional forms, regression families and types of diagnoses?

**4.4 Benefits arising from the research**

4.4.1 Ease of model fitting and comparisons of alternative models

4.4.2 Less confusion and arbitrariness in the fitting of count models

4.4.3 Making sophisticated count modelling available to a wider group of statisticians

4.4.4 Discourage academics from wasteful pursuit of exotic count distributions and encourage them to add more diagnostics to a limited collection of useful and interpretable models

**5. Methodology**

Hilbe [*Negative Binomial Regression*, Cambridge University Press, 2012 ] identifies the following sources of count response models:

Binary: binary logistic and probit regression

Proportional: grouped logistic, grouped complementary loglog

Ordered: ordinal logistic and ordered probit regression

Multinomial: discrete choice logistic regression

Count: Poisson and negative binomial regression

5.1 Identify the most common functional forms used in count regression modelling by re-reading the texts and papers listed in Section 3.

5.2 Identify the modifications to these distributions to improve model fit. Can they be used easily and successfully in mixture modelling? Again by re-reading as above.

5.3 Regression model formulations and parameter estimation methods. Mainly using Cameron & Trividi and Davidian & Giltina texts.

5.4 Read papers on marginal models where the marginal mean is modelled directly in terms of covariates.

5.5 Examine recent work on count transformation models

5.6 Gaps in the process for assessing count models.

5.7 Nature of residuals for count models

5.8 Using residuals to test model assumptions and fit

5. 9 Identifying a successful strategy for identifying, fitting, and evaluating count regression models

**6. Data Sources**

I have identified 5 data sources to use for modelling purposes. The primary response variable is any one of 45 diseases. Potential covariate data has also been obtained, and it is possible that the climatic data might be used to provide counts of exceedances (number of days per month that rainfall or humidity, etc exceeded a specific value) and hence another source of count data for modelling. Details of response data and potential covariates that have already been extracted follow.

**Response Data:**

**6.1 NZ Surveilled Diseases**

The website for Public Health Surveillance is provisioned by The Institute of Environmental Science and Research Ltd (ESR) and is under contract to the Ministry of Health (MOH). The website ( <https://surv.esr.cri.nz> ) states that “This website provides access to selected New Zealand health surveillance data and information”.

Monthly counts for over 30 diseases by DHB are available via the site, e.g. <https://surv.esr.cri.nz/PDF_surveillance/MthSurvRpt/2003/200312DecDHB.pdf>

I downloaded data from September 2001 to September 2020 – a total of 228 months of data.

When I contacted the ESR concerning amalgamation of some of the DHB’s I was informed that these reports are provisional so some of the numbers might differ from finalised figures, presumably available from DHB’s. I have left aside this matter of “measurement error” for follow up at some other time.

As indicated above, the number of diseases is not rigidly fixed from month to month and the DHB’s underwent amalgamation during the reporting period I examined. The amalgamation affected mainly 3 Auckland area DHB’s and hence I have summed the original separate numbers into one Auckland DHB. The total number of diseases encountered in the downloaded data was 45, and the number of DHB’s I amalgamated data to numbered 20. Some of the reported diseases must have been included by mistake, e.g., “Hazardous substances” and “Other”. Fortunately, these were the only two disease name outliers and the remaining diseases has suitably medical-sounding names.

For completeness I include an image of a table of the number of times out of 229 (months) each monthly disease was count was 1 or more at the all-NZ level. 24 out of the 45 diseases had a count of at least 1 for each of the 229 months, at the all-NZ level. The remaining diseases reported counts of at least one on 1, 6, 9, 13, 18, 23, 42, 49, 67, 72, 80, 93, 144, 219, 227 or 228 occasions out of 229 months at the all-NZ level. Hence, at the individual DHB level there appears to be considerable opportunity to experiment with fitting inflation, hurdle, or truncated models. There is also the opportunity to experiment with fitting models at the all-NZ level.

**Covariates:**

**6.2. Monthly Passenger Arrivals:** In order to experiment with the use of covariates in regression models I also downloaded from the Statistics NZ web portal the monthly passenger arrivals NZ by country of residence for the period of interest mentioned previously. Some countries were amalgamated into a Pacific Islands group as their Covid-19 experiences appear similar.

**6.3. Deprivation indices** and their components, by DHB, have also been sourced from Statistics NZ for use when all-NZ modelling is attempted.

**6.4. Housing quality** (mould, dampness) indices by DHB have also been downloaded for potential use when DHB’s are modelled jointly (initially I intend looking only at DHB’s separately).

**6.5. Monthly climatic** summary information (temperature, humidity, number of wet days) has been downloaded from New Zealand's National Climate Database (NIWA - <https://cliflo.niwa.co.nz/> ). To use the multi-site data, I have experimented with obtaining a geospatially distributed subsample of the observation sites around Auckland for summarisation later and inclusion in any Auckland DHB-specific model.

**Appendix A: Frameworks for model fitting and assessment**

<https://acsess.onlinelibrary.wiley.com/doi/full/10.2134/agronj2012.0506>

Nonlinear Regression Models and Applications in Agricultural Research

Sotirios V. Archontoulis, Fernando E. Miguez

Agronomy Journal

Volume107, Issue2

March–April 2015

Pages 786-798

https://doi.org/10.2134/agronj2012.0506

Diagram

Description automatically generated

nlraa: An R package for Nonlinear Regression Applications in Agricultural Research

Fernando Miguez